SIMILARIDAD Y CONTRASTE MEDIANTE PROFUNDIDAD ESTADÍSTICA

Tesis Doctoral

Autor: Ángel López

Director: Juan Romo

Departamento de Estadística

UNIVERSIDAD CARLOS III DE MADRID

Junio 2010

A mis padres, hermanas

y a Carolina

Agradecimientos

En primer lugar, quiero agradecer al Departamento de Estadística de la Universidad Carlos III de Madrid el haberme facilitado todos los recursos necesarios durante el doctorado; al Ministerio de Ciencia e Innovación, al de Educación y a la Comunidad de Madrid, por la financiación recibida a través de los proyectos de investigación ECO2008-05080, SEJ2005-06454, BEC2002-03769 y 06/HSE/0181/2004.

También quiero agradecer a todos aquellos profesores que, desde que comencé la Diplomatura en Estadística en esta universidad, han contribuido en mi formación. De forma muy especial, este agradecimiento está dirigido a los profesores Álvaro Cortínez, Paco Mármol y Juan Romo. A los dos primeros agradezco no sólo su labor docente, sino también la relación personal, ya que, gracias a su motivación, consiguieron que mi interés por la estadística aumentara considerablemente. A Juan le agradezco, como profesor, sus enseñanzas y el interés que mostró por mí, para que hiciera el doctorado; y, como director, sus aportaciones, el apoyo y la atención que me ha brindado estos últimos años. Con él he aprendido a no rendirme cuando los resultados no son los esperados.

Quiero dar las gracias a todos mis amigos y compañeros de departamento, por los buenos momentos que hemos pasado tanto dentro, como fuera de la universidad. En especial a María, que siempre que la he necesitado ha estado ahí para echarme una mano en lo que fuera; a Nacho, quien incluso cuando no tenía tiempo, me lo dedicaba para discutir en *profundidad* mis inquietudes; a Aurora, Lee y Sara, por sus ánimos y por su ayuda en la recta final con los trámites y con el formato de este documento; a Henry y Dalia, por todos los ratos de charla estadística y no estadística que hemos pasado,

tintico en mano; a Javi, Bernardo, Emilio y Mari por su amistad y por ser tan buenos compañeros.

No puedo olvidarme de mis amigos Hugo y Eliud, que siempre me han animado a seguir adelante. Gracias por haber estado siempre ahí y por haber mantenido el contacto a pesar de que yo, en determinados momentos, he estado bastante ausente.

He querido dejar para el final a mi familia, ya que sin ella estoy convencido de que no estaría hoy escribiendo estos agradecimientos.

A mis padres, les agradezco que me hayan animado en el estudio desde niño y también, todo el apoyo que me han ofrecido desde que comencé el doctorado. Éste no se ha limitado a palabras de ánimo. Con su esfuerzo y dedicación, día tras día, han conseguido que mi vida cotidiana haya sido mucho más sencilla, lo que para mi se ha traducido, sobre todo en la última etapa, en un tiempo vital para la finalización de la tesis. A mis hermanas, les agradezco también el apoyo para que siguiera hasta el final y su preocupación por mis estados de ánimo, pero, sobre todo, el hecho de que siempre hayan confiado en mis capacidades. Gracias también a Alejandro, Hugo y Gabriela, quienes, incluso en los momentos en que he estado más decaído, han sido capaces de arrancarme una sonrisa; y a mis cuñados, por todos los momentos agradables y divertidos de nuestras reuniones familiares, que tanto me han ayudado a relajarme y a dejar de lado temporalmente las preocupaciones.

Para mi mujer, Carolina, no tengo suficientes palabras de agradecimiento. Ella ha vivido el día a día de este largo proceso, ha sabido ser paciente con las noches y los fines de semana de duro trabajo, ha sido mi mayor apoyo en los momentos difíciles y mi mejor compañera cuando los resultados han acompañado. Siempre ha estado pendiente de mí y ha hecho todo lo posible para conseguir esos momentos de distracción que tan beneficiosos me han resultado. Sin la estabilidad y el amor que me ha proporcionado, muy probablemente nunca hubiera terminado esta tesis.

Índice general

Capítulo 1

Introducción

Resumen

El objetivo principal de esta tesis doctoral es la exploración de las posibilidades que ofrecen las funciones de profundidad estadística, tanto en la definición de medidas de proximidad en sentido estadístico, como en problemas de clasificación y de contraste de bondad de ajuste. En este capítulo se introduce, en primer lugar, la noción de profundidad estadística. A continuación, se presentan las propiedades deseables de las funciones de profundidad y una clasificación suya según su forma funcional. También se enumeran las funciones más relevantes de la literatura. Por último, se muestran aplicaciones de las profundidades, tanto en el análisis estadístico de datos, como en aplicaciones más complejas.

1.1. Ordenaciones estadísticas de datos

Un área de investigación reciente en estadística es la ordenación de datos en alta dimensión. La estadística multivariante se ha desarrollado ampliamente desde el punto de vista metodológico y teórico desde finales del siglo XIX hasta el punto de conseguir adaptar la metodología unidimensional con gran éxito. Frente a ésta tiene muchas más aplicaciones gracias a la posibilidad de estudiar la relación entre variables, permitiendo así la utilización de una gran variedad de modelos, tanto para datos estáticos como temporales. Pero dentro de este desarrollo existe un problema difícil de resolver que está siendo abordado durante las últimas décadas. Es el problema de la ordenación de datos en alta dimensión.

Parece natural la extensión al caso multivariante de todos los aspectos desarrollados para la estadística univariante. De ahí que el siguiente paso tras la elaboración de técnicas multivariantes, que van desde la descripción de los datos hasta la modelización, estimación y predicción, sea la adaptación de los métodos basados en rangos y ordenaciones de conjuntos de datos. Cuando se estudia una sola variable, la ordenación de un conjunto de datos es trivial: de menor a mayor. Pero cuando la dimensión es mayor, ese concepto es más complejo.

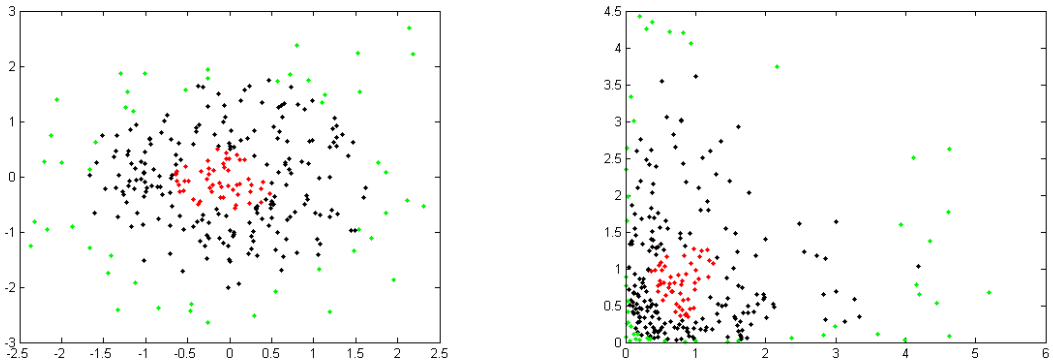
La introducción de los modelos no paramétricos está en parte motivada por los problemas de estimación de los modelos paramétricos en los casos en que se tienen muestras contaminadas; es decir, mayoritariamente las observaciones provienen de un modelo paramétrico o al menos cumpliendo hipótesis como, por ejemplo, la simetría, pero se tiene la presencia de una pequeña proporción de datos cuya distribución, alejada de la del grueso de las observaciones, no es en absoluto de interés. Éstos pueden ocasionar graves errores en la estimación de los parámetros del modelo en caso de no ser eliminado su impacto. Este efecto indeseable ha impulsado el desarrollo de la estimación robusta de parámetros y momentos, entendiéndose por estimador robusto aquel que presenta poca variación ante pequeños cambios en la distribución de los datos. Se puede entender como una forma de continuidad respecto de la distribución poblacional F. Un caso ampliamente estudiado es el de la estimación del centro de una distribución simétrica. Además de la

media y la mediana se han definido varias formas de estimar el centro como, por ejemplo, las medias recortadas, para las que las dos medidas mencionadas son casos límite. Las medidas dirigidas a la estimación robusta del centro están basadas en la capacidad de ordenación de los datos con el fin de no dar peso excesivo a observaciones extremas. De ahí que, como extensión de las técnicas univariantes, cuando se intenta estimar de forma robusta en el análisis multivariante sea esencial la existencia de ordenaciones, al menos de forma parcial, de las observaciones.

En Barnett (1976) se recogen distintos tipos de ordenación de datos multivariantes: orden marginal, orden reducido, orden parcial y orden condicional. El primero consiste en el estudio de las distribuciones marginales de las variables, sobre las que se hace una ordenación univariante. Por otro lado se tiene la ordenación reducida, en la que se asigna a cada punto en $\mathbb{R}^d$ un escalar obtenido mediante la combinación de las coordenadas del punto. Se suele emplear una medida de distancia generalizada $(x-\alpha)'\Gamma^{-1}(x-\alpha)$, donde se encontraría la distancia de Mahalanobis. La ordenación parcial consiste en dividir la muestra en grupos y asignar a todos los puntos de cada grupo un rango. Un método para hacer esta división es el denominado *convex hull peeling*, que consiste en obtener la envolvente convexa de todos los puntos y asignar a los que pertenezcan a ella un valor (por ejemplo, 1). Dichos puntos son eliminados y sobre los restantes se vuelve a obtener la envolvente convexa; a los pertenecientes a ésta se les asigna otro valor y el proceso se repite hasta que todos los puntos están ordenados por capas. Por último está el orden condicional, que se basa en la división en bloques del espacio. Se hacen divisiones a través de todas sus componentes hasta que haya una división que permita ordenar los bloques.

Otro concepto de ordenación, y que será la herramienta sobre la que se fundamentan los desarrollos de esta tesis, es el de profundidad de datos multivariantes. Consiste en la ordenación con respecto a una función de distribución y se basa en una ordenación de dentro hacia fuera, es decir, los puntos más profundos de una nube de puntos son los que se encuentran más próximos al centro de la misma, mientras que los más alejados en sentido estadístico de éstos (los más externos de dicha nube) son los menos profundos. Esta noción da lugar a las llamadas funciones de profundidad que asignan a cada punto un

valor real no negativo y que, dado un punto $x \in \mathbb{R}^d$ y una distribución de probabilidad d-dimensional F, se denotan por $P(x; F)$. Las Figuras 1.1(a) y 1.1(b) muestran gráficamente el objetivo de las funciones de profundidad, que es el de la cuantificación de la centralidad de los puntos teniendo en cuenta la forma de la distribución. Así, por ejemplo, en la Figura 1.1(a), debido a la forma elíptica de la muestra, se consideran profundos o centrales (en rojo) los puntos pertenecientes a las elipses de volumen más pequeño y externos (en verde) a los que no están en las elipses más pequeñas. Si la forma de la nube de puntos no es simétrica, como corresponde a la muestra de la la Figura 1.1(b), parece natural pensar que, en un sentido estadístico, la forma de puntos centrales (en rojo) y externos (en verde) se ajuste a la de la nube de puntos.



(a) Muestra simétrica

(b) Muestra asimétrica

Figura 1.1: *Puntos profundos y externos para muestras de diferente forma.*

Para cada punto y función de distribución, se denota la profundidad poblacional como $P(x; F)$. Para obtener una estimación a través de una muestra se tiene que sustituir F por una estimación razonable F_n, denotando la función de profundidad empírica o muestral como $P_n(x) \equiv P(x, F_n)$. Tras la aplicación de una función de profundidad sobre un conjunto de observaciones, si se ordenan los valores obtenidos de mayor a menor se obtiene una ordenación de dicho conjunto de dentro hacia fuera, desde los más internos hasta los más externos.

1.2. Propiedades de las funciones de profundidad

Algunas propiedades de las funciones de profundidad se definen en torno al centro de simetría de la función de distribución. La idea de simetría en $\mathbb{R}^d$ no es única. Existen varias nociones de simetría que pueden ordenarse según lo restrictivas que sean. A continuación se enumeran algunas de ellas:

a. *Simetría esférica.* La distribución de la variable aleatoria X se dice que presenta simetría esférica en torno al punto c si cualquier rotación de $(X - c)$ tiene la misma distribución que $(X - c)$, es decir, que para cualquier matriz ortogonal U la variable $U(X - c)$ posee la misma distribución que $(X - c)$.

b. *Simetría elíptica.* La función de distribución de X es elípticamente simétrica en torno a c si existe una matriz no singular V que transforme la variable X en una distribución esféricamente simétrica, es decir, que VX sea esféricamente simétrica sobre c.

c. *Simetría antipodal.* La distribución de X es antipodalmente simétrica sobre c si la distribución de $(X - c)$ y $-(X - c)$ son idénticas.

d. *Simetría angular.* La distribución de X es angularmente simétrica sobre c si las distribuciones de $(X - c) / \|X - c\|$ y $-(X - c) / \|X - c\|$ condicionadas a que $X \neq c$ son idénticas.

El orden en que se han dispuesto las distintas definiciones concuerda con su restrictividad, es decir, la más restrictiva es la simetría esférica y la menos, la angular. Al punto c se le denomina centro de simetría.

Dos características fundamentales que comparten todas las funciones propuestas en la literatura son que están acotadas y son no negativas, pero además de estas dos propiedades se han establecido otras deseables que han de cumplir para asegurar un comportamiento eficaz. A continuación se enumeran estas cuatro propiedades.

En primer lugar, la función debe tener su máximo en el centro de la distribución, ya que éste es, sin duda, el punto más profundo en sentido estadístico. En el extremo opuesto

nos encontramos con los puntos externos, los menos profundos con valor de la función decayendo cuanto más nos alejemos del centro; debido a la restricción de no negatividad se puede establecer que en el infinito la profundidad sea cero (es posible que algunas funciones deban ser modificadas mediante la sustracción de alguna cantidad).

Otra propiedad importante es la invarianza afín. Si tenemos una ordenación para un conjunto de datos y un valor de profundidad asociada a cada uno, al realizar una transformación afín de los mismos y ordenar el nuevo conjunto de datos, dicha ordenación y dichos valores para cada punto han de ser los mismos. Esta propiedad puede resultar exigente para algunas funciones de profundidad, así que es posible relajarla y exigir simplemente la invarianza: que la ordenación de los puntos sea igual para los datos con y sin transformación.

Por último, se encuentra la propiedad de decrecimiento monótono desde el punto más profundo. Lo que significa que, dado un punto x, la profundidad para todos los puntos del segmento que une el centro con x debe ser monótona decreciente conforme nos alejamos del centro.

Estas propiedades fueron inicialmente estudiadas en Liu (1990) para la función de profundidad simplicial y, posteriormente, y de forma genérica fueron agrupadas en la siguiente definición de profundidad estadística en Zuo y Serfling (2000a).

Definición 1.1 *Sea* $\mathbf{F}$ *la clase de funciones de distribución en* $\mathbb{R}^d$*. Sea la aplicación* $P(\cdot;\cdot) : \mathbb{R}^d \times \mathbf{F} \to \mathbb{R}$ *acotada, no negativa y verificando las siguientes propiedades:*

i) $P(Ax+b; F_{AX+b}) = P(x; F_X)$ *para cualquier vector* X *en* $\mathbb{R}^d$*, cualquier matriz* A *de dimensión* $d \times d$ *no singular y cualquier vector* b *de dimensión* d*, donde* F_X *es la función de distribución del vector aleatorio* X*;*

ii) $P(\theta; F) = \sup_{x \in \mathbb{R}^d} P(x; F)$*, para cualquier* $F \in \mathbf{F}$ *cuyo centro sea* θ*;*

iii) Para cualquier $F \in \mathbf{F}$ *con punto más profundo* θ *y para todo* $\alpha \in [0,1]$*, se cumple que* $P(x; F) \leq P(\theta + \alpha(x-\theta); F)$*; y*

iv) $P(x; F) \to 0$ *cuando* $\|x\| \to \infty$*, para cada* $F \in \mathbf{F}$*.*

Entonces se dice que $P(\cdot; F)$ es una función de profundidad.

1.3. Clasificación de las funciones de profundidad

En Zuo y Serfling (2000a) se proponen cuatro formas funcionales para la construcción de funciones de profundidad y se estudian con respecto a las cuatro propiedades anteriores. El objetivo principal es obtener funciones no negativas, acotadas y que asignen valores altos a puntos centrales de la nube de puntos y bajos a los más externos. A continuación se presentan los cuatro tipos de funciones propuestos en ese trabajo.

1.3.1. Funciones de profundidad de tipo A

Dada una función acotada y no negativa $h(x; x_1, \ldots, x_r)$ que mide la proximidad entre el punto x y los puntos $x_1, x_2, ..., x_r$ (toma valores altos si x está próximo a $x_1, x_2, ..., x_r$), se construye la función promedio de dichas distancias desde el punto x hasta la muestra aleatoria de tamaño r: $P(x; F) = E[h(x; X_1, ..., X_r)]$, donde $X_1, ..., X_r$ es una muestra aleatoria de F.

1.3.2. Funciones de profundidad de tipo B

Dada cualquier función no negativa y no acotada que mida de algún modo la distancia entre x y el conjunto de puntos $x_1, x_2, ..., x_r$, que se denota por $h(x; x_1, ..., x_r)$, se define $P(x; F) = (1 + E[h(x; X_1, ..., X_r)])^{-1}$, donde $X_1, ..., X_r$ es una muestra aleatoria de F.

1.3.3. Funciones de profundidad de tipo C

Se mide la atipicidad del punto x con respecto al centro de la distribución F a través de una función, generalmente no acotada, $O(x; F)$. Su correspondiente función de profundidad acotada es de una forma funcional similar a la anterior $P(x; F) = (1 + O(x; F))^{-1}$.

1.3.4. Funciones de profundidad de tipo D

Sea $\mathbf{C}$ una clase de subconjuntos cerrados de $\mathbb{R}^d$ y μ una medida de probabilidad en $\mathbb{R}^d$. Se define la función de profundidad como

$$P(x;\mu) = \text{ínf}\left\{\mu(C) : x \in C \in \mathbf{C}\right\}.$$

Así se define la profundidad del punto x como la probabilidad más pequeña acumulada por un conjunto $C \in \mathbf{C}$ que contiene al punto x. Existen dos condiciones que debe cumplir la clase $\mathbf{C}$ para que la función sea útil:

i) Si $C \in \mathbf{C}$, entonces $\overline{C^c} \in \mathbf{C}$;

ii) Para $C \in \mathbf{C}$ y $x \in C^o$, existe $C_1 \in \mathbf{C}$ con $x \in \partial C_1$, $C_1 \subset C^o$;

donde $\partial C, C^c, C^o$ y $\overline{C}$ son respectivamente la frontera, el complemento, el interior y la clausura de C.

1.4. Ejemplos de funciones de profundidad

Para cada uno de los cuatro tipos de funciones de la clasificación anterior existen varias definiciones propuestas en la literatura. Éstas atienden a conceptos geométricos, tanto de volumen, como de distancia en sentido estadístico e incluso de atipicidad de las observaciones con respecto al resto. En esta sección se presentan y clasifican algunas de las muchas definiciones propuestas en la literatura.

Profundidad de Mahalanobis: Está basada en la distancia de Mahalanobis (véase Mahalanobis (1936)) de cada punto al vector de medias. La distancia, para funciones de densidad que en alguna dimensión tengan soporte no acotado y no se anulen sobre ésta, no estará acotada (por ejemplo en el caso de la normal), de ahí que se la realice una transformación para conseguir una función de profundidad acotada. Su definición formal es

$$PM(x;F) = \left[1 + (x-\mu_F)' \Sigma_F^{-1} (x-\mu_F)\right]^{-1},$$

donde μ_F es la esperanza de la distribución F y Σ_F su matriz de covarianzas. La distancia de Mahalanobis es no negativa debido a la estructura de la matriz de covarianza y su valor mínimo se alcanza para $x = \mu_F$, obteniéndose en ese punto una profundidad igual a 1. Para puntos alejados de la media en el sentido de esta distancia se obtienen valores de profundidad cercanos a cero, siendo éste el límite para puntos en el infinito. Se tiene que $PM(x; F)$ toma valores entre cero y uno.

Una de sus características principales es la necesidad de existencia de los primeros dos momentos de la distribución. Esta función mide la centralidad mediante elipses, intuyéndose por tanto, la limitación existente para distribuciones no simétricas, no sólo porque las curvas de nivel de las funciones de densidad no sean elípticas, sino también porque está basada en la media muestral del conjunto de datos, que es poco robusta.

La versión muestral se obtiene sustituyendo el vector μ_F y la matriz Σ_F por estimaciones muestrales suyas, es decir, por ejemplo, $PM(x; F) = \left[1 + (x - \overline{x})' S^{-1} (x - \overline{x})\right]^{-1}$, donde $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$ y $S = n^{-1} \sum_{i=1}^{n} (x_i - \overline{x}) (x_i - \overline{x})'$.

Esta profundidad es afín invariante. En Zuo y Serfling (2000a) se estudian las otras tres propiedades mencionadas, concluyéndose que si la función de distribución de los datos es simétrica, cumple las cuatro y, por tanto, es una función de profundidad estadística según la definición.

La profundidad de Mahalanobis se encuadra dentro de las funciones de tipo C, ya que se emplea una medida de atipicidad del punto x :

$$A(x; F) = (x - \mu_F)' \Sigma_F^{-1} (x - \mu_F),$$

resultando

$$P(x; F) = (1 + A(x; F))^{-1} = \left(1 + (x - \mu_F)' \Sigma_F^{-1} (x - \mu_F)\right)^{-1} = PM(x; F).$$

Profundidad semiespacial de Tukey: Introducida por Tukey (1975), a cada punto x se le asigna como profundidad el mínimo de la probabilidad de los semiespacios que contienen al punto x. Para una distribución elípticamente simétrica, si se cálcula la profundidad del centro de la misma, todos los semiespacios que pasan por él dejan a ambos lados una

probabilidad igual a 1/2. Esa será, por tanto, la cota superior para la profundidad del semiespacio.

Esta definición se encuadra dentro de las funciones de tipo D, siendo la familia de subconjuntos sobre los que se halla el ínfimo la de los semiespacios cerrados **H** que verifican las dos propiedades necesarias para una correcta definición:

i) Si $H \in \mathbf{H}$ entonces $\overline{H^c} \in \mathbf{H}$. Lo que es obvio, ya que el complementario de un semiespacio cerrado es el semiespacio abierto con vector director opuesto, y si añadimos la frontera se convierte en cerrado.

ii) Dado un semiespacio H y un punto x de su interior, entonces existe otro semiespacio H' contenido en el primero y en cuya frontera está el punto x. Se cumple también, pues sólo habría que trasladar el semiespacio anterior sobre el punto x.

Dada la función de distribución d-dimensional F, la profundidad semiespacial se define como

$$PSem(x;F) = \text{ínf}\left\{Pr(H) : H \textit{ semiespacio cerrado},\ x \in H\right\}.$$

La versión muestral de esta profundidad se obtiene a partir de una estimación de la función de distribución, por ejemplo, mediante la función empírica: $PSem(x;F_n)$. Esta función de profundidad verifica las cuatro propiedades de la definición (véase Zuo y Serfling (2000a)).

En Donoho y Gasko (1992) se estudian detalladamente propiedades asintóticas de esta profundidad y del punto más profundo, realizando además un estudio acerca de los puntos de ruptura de estimadores basados en la ordenación por esta profundidad. Más propiedades asintóticas pueden encontrarse en Bai y He (1999) y Massé (2002). En Yeh y Singh (1997) pueden encontrarse desarrollos sobre las regiones de confianza.

En cuanto a las propiedades asintóticas de la región recortada de nivel α (véase Nolan (1992)), definida como la intersección de los semiespacios que dejan al menos una probabilidad $(1-\alpha)$, se tiene que, bajo ciertas condiciones de unicidad del semiespacio mínimo, dicha región es un conjunto convexo y consistente con la región recortada de la medida de probabilidad μ; es decir, dado un punto x, el semiespacio muestral con probabilidad

menor o igual que α es consistente con el semiespacio distribucional. Basándose en estas regiones recortadas es posible la obtención de medidas de localización robustas.

Convex hull peeling: Introducida en Barnett (1976). Consiste en la elaboración iterativa de capas basadas en la envolvente convexa del conjunto de datos. A cada observación dentro de la misma capa se le asigna el mismo valor, creando así clases de equivalencia. El número de capas depende del de observaciones y de la dimensión del problema, y además está acotado. Para problemas en $\mathbb{R}^d$ se necesitan al menos $d+1$ puntos distintos para obtener una envolvente convexa de dichos puntos. Así que el mayor número de capas que se va a obtener será $\lfloor \frac{n}{d+1} + 1 \rfloor$ donde $\lfloor \cdot \rfloor$ denota la parte entera. El proceso iterativo es el siguiente:

1) Crear el conjunto de índices $I = 1, ..., n$. Hacer $contcapa = 1$.

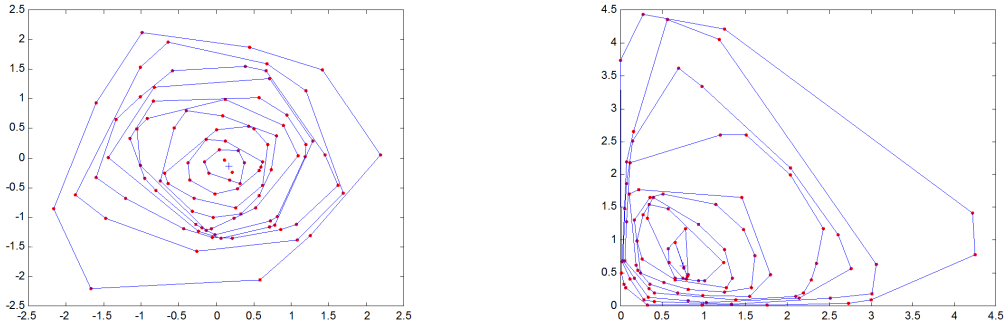
2) Hallar la envolvente convexa de todos los puntos de I. Asignar el valor $contcapa$ a los vertices de la envolvente, cuyos índices se denotan con $I_{contcapa}$. Eliminar dichos puntos del conjunto: $I = I \backslash I_{contcapa}$.

3) Si $\#\{I\} > d+1$: actualizar $contcapa = contcapa + 1$. Volver a 1; si no, terminar el proceso asignando los puntos restantes a la última capa.

Como se puede observar, el proceso acaba, bien cuando quedan por asignar $d+1$ puntos que formarán la última envolvente, o bien cuando sobre un número menor y no se pueda formar una envolvente. Cuando se da el caso en que la última capa está formada por más de un punto, se define el más profundo como la media de todos ellos.

Esta profundidad presenta el problema de que no tiene definición poblacional, con el consiguiente problema de inferencia sobre las capas obtenidas para la muestra. Además, la ordenación de las observaciones se realiza por clases que pueden ser muy numerosas (al menos $d+1$ puntos). Las figuras 1.2(a) y 1.2(b) muestran las envolventes para las muestras normal y exponencial, y también el punto más profundo. Se observa que en ambos casos la estimación del centro es bastante correcta.

Debido a que no posee definición poblacional no es posible clasificar el método de la envolvente convexa en ninguno de los tipos definidos en el apartado anterior.



(a) Distribución normal (b) Distribución exponencial

Figura 1.2: *Punto más profundo según envolvente convexa.*

Profundidad de Oja: Introducida en Oja (1983). Se basa en la formación de envolventes convexas de $d+1$ puntos, es decir, símplices, para los que se calcula su volumen. Recoge la información del punto x al introducirlo en todos los símplices. Dados x y $X_1, X_2, ..., X_d$ (muestra aleatoria de F de tamaño d) se define el simplex $S[x, X_1, X_2, ..., X_d]$ y se calcula su volumen. Si el volumen es grande significa que el punto x está alejado de los otros d puntos y si es pequeño que está cerca del hiperplano definido por ellos. Así pues, nos encontramos ante una función que mide distancia entre puntos: $h(x, X_1, X_2, ..., X_d) = Vol(S[x, X_1, X_2, ..., X_d])$ y que es no acotada. La función de profundidad de Oja se define como

$$PO(x; F) = \left[1 + E_F\left(Vol(S[x, X_1, X_2, ..., X_d])\right)\right]^{-1},$$

perteneciendo por tanto a las funciones de tipo B.

Su versión muestral se obtiene a partir del volumen medio para todas combinaciones posibles de d puntos de la muestra, es decir,

$$PO_n(x) = \left[1 + \binom{n}{d}^{-1} \sum_{1 \le i_1 < ... < i_d \le n} Vol(S[x, x_{i_1}, x_{i_2}, ..., x_{i_d}])\right]^{-1}.$$

Para esta profundidad no se cumple la propiedad de invarianza afín. Existen variantes de la función de profundidad de Oja que sí cumplen esa propiedad. La función así definida sí conserva la ordenación, que no el valor asignado (ya que se trata de volúmenes sin estandarizar de ningún modo), ante transformaciones lineales.

Profundidad simplicial: Función de profundidad definida en Liu (1990). La profun-

didad para el punto x se define como la probabilidad de que pertenezca a un símplex formado por una muestra de la distribución F de tamaño $d+1$. En otras palabras, dada dicha muestra se comprueba si el punto x es combinación lineal convexa de estas observaciones. Así pues, se espera que un punto externo de la distribución pertenezca a una pequeña proporción de símplices, ya que para que esté en un símplex debe haber un punto en la muestra más alejado que él y, por tanto, menos probable de encontrar. Sin embargo, para un punto central se espera que haya un gran número de puntos que al formar el símplex le contenga.

Así, se tiene $PS(x;F) = Pr\{x \in S[X_1, X_2, \ldots, X_{d+1}]\} = E\left[I_{S[X_1,X_2,\ldots,X_{d+1}]}(x)\right]$, donde $I_{S[X_1,X_2,\ldots,X_{d+1}]}(x)$ es la función indicadora que se define como

$$I_{S[X_1,X_2,\ldots,X_{d+1}]}(x) = \begin{cases} 1 & \text{si } x \in S[X_1, X_2, \ldots, X_{d+1}] \\ 0 & \text{en otro caso.} \end{cases}$$

Denotando esta función como $h(x, X_1, X_2, \ldots, X_{d+1}) = I_{S[X_1,X_2,\ldots,X_{d+1}]}(x)$ es posible clasificar la profundidad simplicial dentro de las funciones de tipo A.

En la versión muestral la suma se realiza a través de todos los posible símplices,

$$PS_n(x) = \binom{n}{d+1}^{-1} \sum_{1 \le i_1 < \ldots < i_{d+1} \le n} I(x \in S[x, x_{i_1}, x_{i_2}, \ldots, x_{i_{d+1}}]).$$

Estudios sobre su comportamiento como función de profundidad en el sentido de la definición se encuentran en Liu (1990) y Zuo y Serfling (2000a), donde se establece la invarianza afín de la función y se asegura el cumplimiento de todas las propiedades para funciones de distribución angularmente simétricas. Su peor comportamiento se encuentra en distribuciones discretas, donde ni la propiedad de maximalidad ni la de monotonía están aseguradas.

En Liu (1990) se prueba por un lado que, para funciones de distribución absolutamente continuas con densidades acotadas, $PS_n(\cdot)$ es uniformemente consistente:

$$\sup_{x \in \mathbb{R}^d} |PS_n(x) - PS(x;F)| \xrightarrow[n\to\infty]{c.s} 0,$$

y, por otro, la convergencia casi segura del punto más profundo de la muestra $\hat{\mu}$ al punto μ, si éste es el único máximo de $PS(\cdot;F)$.

Además, en Dümbgen (1992) se estudian teoremas límite para la profundidad simplicial y se asegura que bajo ciertas hipótesis sobre F, si la sucesión F_n converge débilmente a F, entonces $\|PS_n(\cdot) - PS(\cdot; F)\|_\infty \to 0$. También establece la normalidad asintótica de los L-estadísticos formados a partir de funciones de pesos diferenciables aplicadas sobre la profundidad de cada punto del espacio.

Profundidad de la mayoría: Singh (1991). Guarda similitudes con respecto a la profundidad de Tukey, ya que consiste también en la elaboración de semiespacios o, más concretamente, en la construcción de hiperplanos, para los que se comprueba si el punto x pertenece a la cara mayoritaria o no; es decir, dado un punto x, se toma una muestra de tamaño d y se calcula el hiperplano que pasa por dichos puntos; posteriormente se verifica si el punto x pertenece al semiespacio que contiene más de la mitad de la probabilidad. La profundidad se define formalmente como la probabilidad de que el punto x pertenezca a la cara mayoritaria obtenida por el hiperplano que pasa por los puntos de una muestra aleatoria de tamaño d

$$PMJ(x; F) = P\left\{ \begin{array}{c} x \in \text{ a la cara mayoritaria del hiperplano} \\ \text{formado por } (X_1, X_2, \ldots, X_d) \end{array} \right\},$$

donde $(X_1, X_2, \ldots, X_d)$ es una muestra aleatoria de tamaño d. La función de profundidad pertenece a las de clase A, donde $r = d$ y la función $h(x; x_1, \ldots, x_d) = I\left\{x \in H^F_{x_1, x_2, \ldots, x_d}\right\}$, $x \in \mathbb{R}^d$, donde $H^F_{x_1, x_2, \ldots, x_d}$ es el semiespacio acotado por el hiperplano formado por los puntos $x_1, x_2, \ldots, x_d$ que contiene más de la mitad de las observaciones.

Profundidad por verosimilitud: Esta profundidad, tal como se ha definido en Fraiman y Meloche (1999), se corresponde con la función de densidad, es decir, $PV(x; F) = f(x)$ cuya versión muestral puede obtenerse mediante una estimación de densidad kernel, con el consiguiente problema de elección del ancho de banda. Además, no verifica la invarianza afín.

Profundidad L^p: Se define como

$$PL^p(x; F) = \left(1 + E\left[\|x - X\|_p\right]\right)^{-1}.$$

Pertenece a la clase B de funciones, ya que mide distancias mediante la norma $\|\cdot\|_p$ es decir, dada $h(x; x_1) = \|x - x_1\|_p$.

Profundidad por proyecciones: Mide la atipicidad de cada punto utilizando una función $A(x;F)$ que consiste en obtener la mayor discrepancia de la proyección del punto x sobre direcciones unitarias u con respecto a la mediana de la proyección del vector aleatorio X sobre dicha dirección

$$A(x;F) = \sup_{\|u\|=1} \frac{|u'x - Med(u'X)|}{MEDA(u'X)},$$

donde $u'x$ es el producto vectorial $\langle u, x\rangle$, el vector X tiene distribución F, Med representa la mediana y MEDA representa la mediana de los valores absolutos de las desviaciones respecto a la mediana. La función de profundidad se define como

$$PP(x;F) = (1 + A(x;F))^{-1}.$$

Sus propiedades de convergencia y puntos de ruptura han sido estudiados en Zuo (2003).

Profundidades angulares: En Liu y Singh (1992) se extienden las profundidades simplicial y de Tukey, y se define la de la distancia del arco para observaciones angulares (sobre esferas) y se estudian sus propiedades y sus puntos de ruptura.

Profundidad del Zonoide: En Koshevoy y Mosler (1997) se define otra noción de profundidad basada en las regiones (recortadas) del zonoide, para las que estudian sus propiedades.

Profundidad asociada a la mediana L_1: Definida en Vardi y Zhang (2000). Dada una definición de mediana multivariante θ y una función de distribución d-dimensional F, para la obtención de la profundidad de un punto x en $\mathbb{R}^d$, se calcula la mínima masa probabilística ω necesaria para hacer de x la mediana multivariante de la mixtura $(\omega\delta_x + F)/(\omega + 1)$, donde δ_x es la función de distribución de una variable aleatoria degenerada en el punto x. Más formalmente se define como:

$$PL_1 = 1 - \text{ínf}\left\{\omega \geq 0 : \theta\left(\frac{\omega\delta_x + F}{1+\omega}\right) = x\right\}.$$

Esta profundidad es computacionalmente menos exigente que muchas de las introducidas previamente. Un algoritmo para su cálculo puede encontrarse en Vardi y Zhang (2000). En Jornsten et al. (2002) y Jornsten (2004) se aplica esta profundidad en métodos de búsqueda de conglomerados y clasificación.

Profundidad por bandas: Esta profundidad fue introducida en López-Pintado y Romo (2009). Debido a su reducido coste computacional, puede ser aplicada en datos de alta dimensión. La idea de esta función consite en el cálculo de la probabilidad de que el punto x esté contenido en bandas aleatorias formadas por b puntos. Pertenece, por tanto, al grupo de funciones tipo A. La banda aleatoria de b puntos se define como

$$B\left(x_1, x_2, \ldots, x_b\right) = \left\{ y \in \mathbb{R}^d : \forall k \in \{1, 2, \ldots, d\}, \min_{i=1,\ldots,b} x_i^{k)} \leq y^{k)} \leq \max_{i=1,\ldots,b} x_i^{k)} \right\},$$

donde $y^{k)}$ es la coordenada k-ésima del vector y y $x_i^{k)}$ es la coordenada i-ésima del punto x_i. La componente de la profundidad por bandas de un punto x en $\mathbb{R}^d$ con respecto a una función de distribución F para bandas formadas por b puntos se define como

$$\begin{aligned} PB^{b)}\left(x; F\right) &= Pr\left[x \in B\left(X_1, X_2, \ldots, X_b\right)\right] \\ &= E\left[\prod_{k=1}^{d} I\left\{ \min_{i=1,\ldots,b} X_i^{k)} \leq x^{k)} \leq \max_{i=1,\ldots,b} X_i^{k)} \right\}\right], \end{aligned}$$

donde $x^{k)}$ es la coordenada k-ésima de x, $X_i^{k)}$ la coordenada k-ésima de la variable aleatoria X_i y X_i, con $i = 1, 2, \ldots, b$, son variables aleatorias independientes e idénticamente distribuidas según F. Haciendo la suma de esta componente para bandas formadas por un máximo de B puntos se obtiene la profundidad por B bandas del punto x:

$$PB\left(x; F, B\right) = \sum_{b=2}^{B} PB^{b)}\left(x; F\right), \ B \geq 2.$$

Su versión muestral queda definida como

$$PB_n\left(x; B\right) = \sum_{b=2}^{B} \binom{n}{b}^{-1} \sum_{(i_1, i_2, \ldots, i_b) \in J_b} \prod_{k=1}^{d} I\left\{ \min_{i \in \{i_1, i_2, \ldots, i_b\}} x_i^{k)} \leq x^{k)} \leq \max_{i \in \{i_1, i_2, \ldots, i_b\}} x_i^{k)} \right\},$$

donde el conjunto de índices J_b es igual a $\{i_1, i_2, \ldots, i_b : 1 \leq i_1 < i_2 < \ldots < i_b \leq n\}$.

Profundidad por bandas modificada: Debido a que en la práctica la aplicación de la profundidad por bandas puede ser demasiado restrictiva en alta dimensión, en López-Pintado y Romo (2009) se define también una versión modificada en la que no se exige la pertenencia de todas las coordenadas del punto sobre el que se calcula la profundidad, sino que se mide el número de coordenadas en que el punto sí está dentro de las bandas.

Más formalmente, para bandas formadas por un número b de puntos de $\mathbb{R}^d$, dado el punto x en $\mathbb{R}^d$ y una función de distribución F, se denota por $PBM^{b)}(x;F)$ a la cantidad que representa el porcentaje medio de coordenadas para las que el punto x está dentro de bandas aleatorias formadas por b puntos, es decir,

$$\begin{aligned} PBM^{b)}(x;F) &= E\left[\frac{1}{d}\sum_{k=1}^{d} I\left\{\min_{i\in\{1,2,\ldots,b\}} X_i^{k)} \leq x^{k)} \leq \max_{i\in\{1,2,\ldots,b\}} X_i^{k)}\right\}\right] \\ &= \frac{1}{d}\sum_{k=1}^{d} E\left[I\left\{\min_{i\in\{1,2,\ldots,b\}} X_i^{k)} \leq x^{k)} \leq \max_{i\in\{1,2,\ldots,b\}} X_i^{k)}\right\}\right] \\ &= \frac{1}{d}\sum_{k=1}^{d} Pr\left[\min_{i\in\{1,2,\ldots,b\}} X_i^{k)} \leq x^{k)} \leq \max_{i\in\{1,2,\ldots,b\}} X_i^{k)}\right] \end{aligned}$$

donde $x^{k)}$ es la coordenada k-ésima de x, $X_i^{k)}$ la coordenada k-ésima de la variable aleatoria X_i y X_i, con $i = 1, 2, \ldots, b$, son variables aleatorias independientes e idénticamente distribuidas según F. La profundidad por bandas modificada queda definida como

$$PBM(x;F,B) = \sum_{b=2}^{B} PBM^{b)}(x;F), \ B \geq 2.$$

Su versión muestral se obtiene promediando entre todas las posibles bandas formadas por un máximo de B puntos de la muestra, es decir,

$$PBM_n(x;B) = \sum_{b=2}^{B} PBM_n^{b)}(x), \ B \geq 2,$$

donde

$$PBM_n^{b)}(x) = \binom{n}{b}^{-1} \sum_{(i_1,i_2,\ldots,i_b)\in J_b} \frac{1}{d}\sum_{k=1}^{d} I\left\{\min_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)} \leq x^{k)} \leq \max_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)}\right\},$$

siendo J_b el conjunto de todas las combinaciones de n índices tomados de b en b.

El principal problema subyacente en la mayoría de las definiciones de profundidad mencionadas es el cálculo de las mismas. El orden de operaciones de los algoritmos intuitivos que se desprenden de cada definición crece exponencialmente con la dimensión del problema. Algoritmos aproximados y exactos para el cálculo de algunas profundidades en dimensión baja pueden encontrarse en Rousseeuw y Ruts (1996), Rousseeuw y Struyf (1998), Vardi y Zhang (2000) y Miller et al. (2003).

1.5. Análisis estadístico de datos mediante la profundidad

En el caso univariante, la ordenación permite tanto la identificación de las observaciones más extremas como la estimación de parámetros de forma más robusta aunque menos eficiente que la estimación usual por máxima verosimilitud. Un ejemplo de estimación robusta de la centralización o localización de una distribución consiste en el método de las medias recortadas, en el que la influencia de los valores extremos se elimina con el fin de que su atipicidad no afecte a la estimación del parámetro de interés. Es viable emplear esta forma de trabajo no sólo para la centralización sino también para la estimación de otros momentos de las distribuciones de probabilidad. De ahí que conseguir las ordenaciones no triviales en el espacio multidimensional sea útil para la construcción de medidas que pretendan ser robustas. También es posible definir nuevas funciones de las observaciones y nuevos gráficos de utilidad tanto para la comparación de muestras como para tratar aspectos tales como la localización, dispersión, asimetría o curtosis.

En Liu et al. (1999) se puede encontrar un amplísimo estudio del análisis de datos multivariantes tras la ordenación a partir de las funciones de profundidad estadística. Se incluyen, además de las técnicas comentadas arriba, métodos inferenciales como la estimación de la matriz de covarianzas y la diagnosis de normalidad.

1.5.1. Definiciones de conjuntos elementales

Antes de realizar cualquier comentario acerca de las técnicas aplicables a un conjunto de datos ordenado, se definen varios conjuntos que están basados principalmente en las curvas de nivel de la función de profundidad con la que se ordenan las observaciones.

Definición 1.2 *El contorno de profundidad t es* $\left\{x \in \mathbb{R}^d : P\left(x; F\right) = t\right\}$.

Definición 1.3 *El conjunto de puntos encerrado por el contorno de profundidad t es* $R\left(t\right) = \left\{x \in \mathbb{R}^d : P\left(x; F\right) > t\right\}$.

Definición 1.4 $C_p = \bigcap\limits_t \left\{R(t) : Pr_F\left(R(t)\right) \geq p\right\}$ *es la región central p-ésima.*

El primero de los conjuntos define la curva de nivel de la función de profundidad asociada al valor t. El segundo, $R(t)$, hace referencia al conjunto de puntos con profundidad mayor que t: como las ordenaciones son desde el centro hacia fuera, son los puntos contenidos por la curva de nivel o contorno de profundidad t. Por último tenemos el conjunto C_p que es la región más pequeña encerrada por todos los contornos que acumulan como mínimo una probabilidad p. A la frontera de dicho conjunto se le denomina curva de nivel *p-ésimo* y se denota por $Q(p)$ o $Q_F(p)$.

Si F es absolutamente continua y f es no nula entonces se tiene que $C_p = R(t_p)$ donde t_p es tal que $Pr_F\left(x \in \mathbb{R}^d : P(x;F) \geq t_p\right) = Pr_F\left(R(t_p)\right) = p$. Por conveniencia computacional la estimación muestral de C_p no se realiza sustituyendo donde corresponda F por F_n y $P(\cdot;F)$ por $P_n(\cdot)$, sino que se obtiene la envolvente convexa de los $\lceil np \rceil$ puntos más profundos, es decir,

$$C_{n,p} = envolvente\ convexa\left(X_{[1]}, X_{[2]}, \ldots, X_{[\lceil np \rceil]}\right),$$

donde $\lceil np \rceil$ es 1 más la parte entera de np si np no es entero, y np si np es entero. A esta estimación se la denomina *p-ésima* envolvente central y a su frontera $Q_n(p)$, curva muestral de nivel *p-ésimo* o superficie empírica del *p-ésimo* cuantil "dentro-fuera". Los empates en la profundidad definen una clase de equivalencia y, a efectos del cálculo de los conjuntos anteriores, ha de tomarse con todos sus elementos.

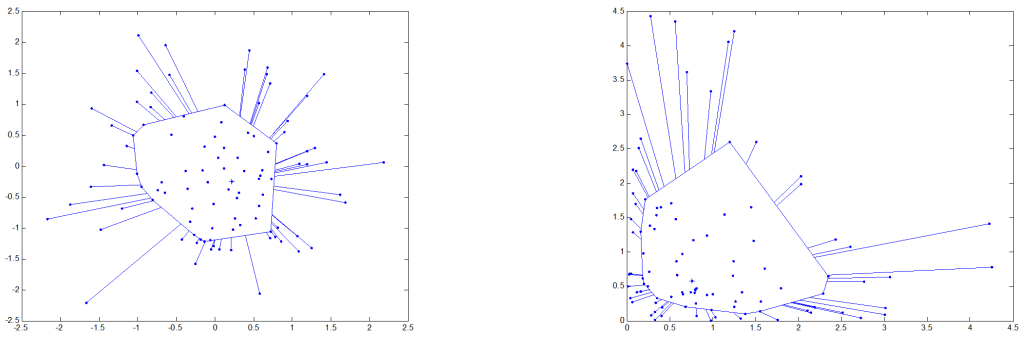
La convergencia de los contornos de profundidad ha sido estudiada en diversos artículos como He y Wang (1997), Nolan (1992), Massé y Theodorescu (1994) y Zuo y Serfling (2000b). Muchos de ellos se basan en los resultados de convergencia de U-procesos de Arcones y Giné (1993) y Arcones et al. (1994).

1.5.2. Métodos gráficos

En cuanto a las posibilidades gráficas que ofrecen las ordenaciones por profundidad, existen numerosas opciones basadas principalmente en diagramas de puntos y en curvas, dependiendo del análisis que se desee llevar a cabo. De todas las opciones posibles, a continuación se introducen tres de las más interesantes. La primera es una extensión al

caso bidimensional del diagrama de caja. La segunda es la representación de curvas que resumen características de los momentos de la distribución, como, por ejemplo, la curva de volumen o curva de escala, que se empleará en el cuarto capítulo para la construcción de un test de bondad de ajuste. Y por último, los denominados *dd*-plot, que consisten en diagramas de dispersión de valores de profundidad sobre distintas muestras.

La extensión del diagrama de caja a dimensión dos se realiza a partir del contorno muestral de nivel 0.5, que se podría interpretar como el rango intercuartílico debido a la ordenación dentro-fuera. Así, se forma la envolvente de la mitad de los puntos más profundos y el resto de los puntos se unen por una línea con el centro (para más detalle véanse Liu et al. (1999) y Rousseeuw et al. (1999), que completa el gráfico con otro contorno). Las figuras 1.3(a) y 1.3(b) muestran, respectivamente, el diagrama de caja para una muestra de distribución normal estándar y para un vector de dos distribuciones exponenciales independientes de media uno, con ordenaciones realizadas según la profundidad del semiespacio.



(a) Distribución normal (b) Distribución exponencial

Figura 1.3: *Diagramas de caja bidimensionales para la profundidad de Tukey.*

Con la ordenación obtenida a través de la profundidad se pueden definir matrices de dispersión muestral recortadas (o ponderadas, asignando un peso a la aportación de cada observación o clase de equivalencia), posibilitando de este modo el estudio de la dispersión a través de un escalar por medio del determinante de estas matrices, empleando los conceptos de varianza generalizada o también de varianza efectiva. Pero, gracias a la ordenación multivariante, es posible calcular para distintos valores de p el volumen de

la *p-ésima* región central, es decir, la que contiene la proporción p de los puntos más profundos. De este modo se obtiene una función de dispersión para $p \in [0,1]$, que se denomina curva de volumen o curva de escala.

La versión muestral puede estimarse de la siguiente forma: para distintos valores de p se seleccionan los $\lceil np \rceil$ puntos más profundos (incluyendo todos los pertenecientes a la clase frontera de p) y calculando el volumen de la envolvente convexa de todos los puntos. De este modo la función teórica y muestral serían, respectivamente,

$$S(p) = Vol(C_p) \;\; y \; S_n(p) = Vol(C_{n,p}).$$

A través de estas dos curvas se puede comparar distribuciones, teniendo en cuenta que, si la curva de una de ellas crece más rápidamente que la otra, tendrá una dispersión mayor. Igual que si una de las curvas está para la mayoría de los puntos por encima de la otra. Por otro lado, el volumen de la envolvente convexa para algún valor fijo como por ejemplo 0.5 sería el equivalente del rango intercuartílico, con lo que se estudiaría la dispersión para la mitad de puntos más profunda. Parece lógico pensar que para una familia de distribuciones con varianzas proporcionales (véase la figura 1.4, que contiene las curvas para dos muestras de normales con varianzas proporcionales), una de ellas deberá estar sistemáticamente por encima de la otra y ambas deberían presentar patrones de crecimiento similares. Esto motiva su aplicación en uno de los contrastes de bondad de ajuste del Capítulo 4.

La utilidad de la función volumen no se limita a la comparación de muestras a través de los gráficos anteriores. Como se muestra en Liu et al. (1999), es posible emplear dicha función unida a algún método de remuestreo, para la observación de la eficiencia de distintos estimadores para descubrir si de forma sistemática un estimador se comporta mejor que el resto. En ese mismo trabajo aparecen también otros métodos gráficos basados en curvas para determinar si un conjunto de observaciones presenta simetría o no (simetrías esférica, elíptica, antipodal y angular) e incluso métodos para la determinación de la curtosis mediante la aplicación de la curva de Lorenz (Lorenz (1905)).

Se concluye esta sección de métodos gráficos con los *dd*-plot, que consisten en diagramas de dispersión de los valores de profundidad de conjuntos de puntos. Su equivalente

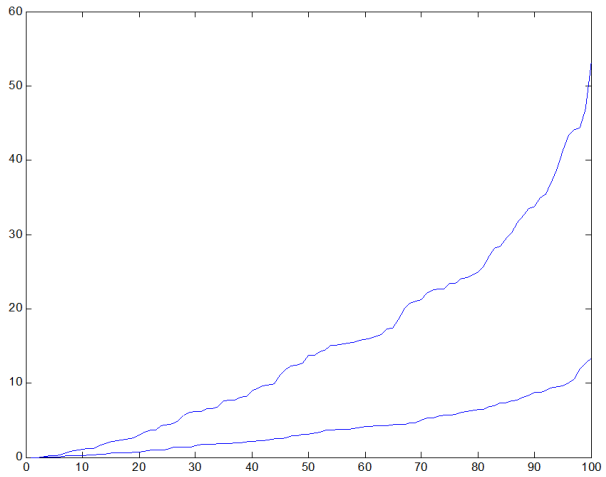


Figura 1.4: *Curvas de volumen de dos muestras normales con matrices de covarianzas proporcionales.*

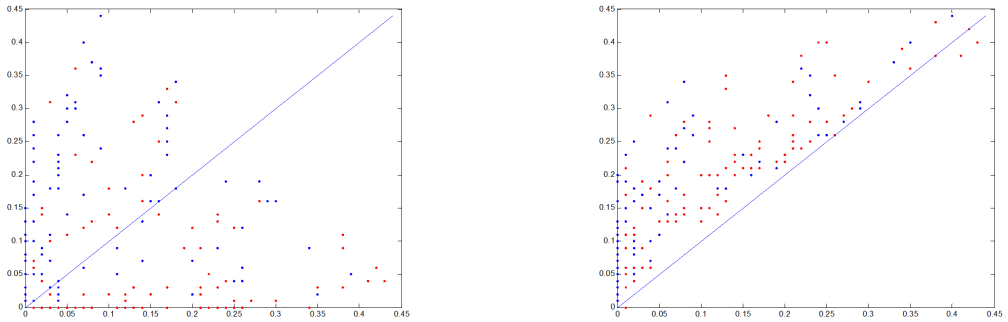
univariante se correspondería con los gráficos cuantil-cuantil. Son de utilidad para realizar comparaciones entre muestras, ya que con ellos se pueden diagnosticar cambios en la localización, en la escala y en la forma con sencillos diagramas bidimensionales. Las comparaciones que pueden llevarse a cabo son para comparar distribuciones entre sí, muestras con distribuciones y muestras entre sí. En este último escenario se desconoce la distribución de las variables a comparar (X e Y) y sólo se dispone de sus muestras, $\mathbb{X} = \{x_1, x_2, \ldots, x_n\}$ e $\mathbb{Y} = \{y_1, y_2, \ldots, y_m\}$. A partir de las funciones empíricas se calculan las profundidades de la muestra formada por $\mathbb{X} \cup \mathbb{Y}$. La combinación de muestras se realiza con el objetivo de representar las posibles direcciones desde el centro hacia fuera, ya que un valor de profundidad genera una región p-ésima y hay que tener en cuenta de algún modo sobre qué punto de la frontera de esa región se está; es decir, no importa sólo el valor de la profundidad en un punto, también interesa saber cuál es la profundidad de ese punto con respecto a las dos distribuciones empíricas.

De esta manera se construye el siguiente conjunto

$$DD\left(F_n, G_m\right) = \left\{\left(P\left(x; F_n\right), P\left(x; G_m\right)\right) \ \textit{para todo}\ x \in \mathbb{X} \cup \mathbb{Y}\right\},$$

cuyo gráfico será utilizado para comprobar si ambas muestras provienen de la misma

distribución o no. Si ambas muestras proceden de la misma distribución se espera que los puntos se concentren en torno a la bisectriz del primer cuadrante. Si esto no sucede, ya sea porque no comparten la misma familia de distribución o bien porque sus centros o variabilidades son diferentes, los puntos estarán alejados de la bisectriz y presentando patrones particulares que puedan denotar sus diferencias. Por ejemplo, las Figuras 1.5(a) y 1.5(b) muestran, respectivamente, los *dd*-plot para dos muestras procedentes de la misma familia de distribuciones con diferencias en la media (1.5(a)) o en la dispersión (1.5(b)).



(a) Cambio en la media

(b) Cambio en la varianza

Figura 1.5: *dd-plot para muestras de la misma distribución con cambios en los parámetros.*

1.5.3. Localización

Existen diferentes tipos de estimadores de localización en el caso univariante, pero si el objetivo es realizar una estimación paramétrica de manera robusta, no es aconsejable el uso de alguno de ellos, tales como máxima verosimilitud, ya que para el supuesto de normalidad, éste se corresponde con la media muestral y, como es sabido, es fácilmente maleable, es decir, su punto de ruptura es cero. En Bickel y Lehmann (1975) y Huber (1972), se hace una recopilación de estimadores robustos como los L-estadísticos (combinaciones lineales de estadísticos de orden), los R-estadísticos (basados en rangos) y los M-estadísticos.

En alta dimensión, el estimador más natural del centro para cada una de las definiciones de profundidad es el punto más profundo. Parece lógico pensar además que, si

ha habido algún empate dentro de los puntos más profundos, haya que tomar la media muestral de los mismos como estimador del centro, ya que al obtener un empate no hay información suficiente para seleccionar a cualquiera de ellos.

Como se ha venido observando a través de los análisis previos, existe la posibilidad de que haya diferencias entre métodos tanto en la ordenación, como en la elección del punto más profundo. Estas diferencias se han comprobado al aplicar los métodos sobre muestras cuya forma distribucional era diferente. Con esos ejemplos se ha desechado el uso de la profundidad de Mahalanobis como estimador de mediana multivariante, salvo para distribuciones con curvas de nivel elípticas. Pero si bien el punto más profundo parece un buen estimador de mediana multivariante en el sentido de robustez, posiblemente esté lejos de ser eficiente. Por lo tanto, hay que intentar encontrar un compromiso entre ambas propiedades. En este sentido se pueden extender los conceptos de los L-estadísticos univariantes, que consisten en una ponderación de los elementos muestrales de forma que sea posible eliminar la influencia de los puntos más externos que puedan tener un elevado índice de atipicidad.

Por otro lado, cuando se dispone de una muestra y se desea calcular la mediana, en general (salvo para la profundidad de Mahalanobis), sólo unos pocos datos influyen en su valor final, de ahí que existan un gran número de estimadores que, al asignar pesos de acuerdo con la ordenación obtenida, ofrezcan un concepto de localización que se sitúa entre la media de todas las observaciones y la mediana como el punto más profundo; por ejemplo, el caso de las medias recortadas.

En alta dimensión se obtiene además una dirección de asimetría en caso de que la distribución no sea simétrica. Cuando se trabaja en una dimensión suele describirse la variable aleatoria de interés como asimétrica positiva o negativa, según sea la cola más pesada hacia la derecha o la izquierda, respectivamente. Sin embargo, cuando se estudian variables d-dimensionales no es viable verificar la asimetría de dicha forma, salvo si el estudio se hace por componentes, forma que se ha desechado en la ordenación desde un primer momento. En cambio, es posible tomar como dirección de asimetría el vector diferencia entre media y centro muestrales (poblacionales), ofreciendo así una dirección

y sentido sobre la cual la cola es más pesada que sobre la misma dirección y sentido opuesto.

Para definir los *L*-estadísticos basados en la idea de profundidad (véase Liu et al. (1999)), en adelante *PL*-estadísticos, es necesario primero la definición de un proceso aleatorio basado en los estadísticos de orden por profundidad y de una función de pesos que se integrará con el proceso anterior para obtener dichos estadísticos.

Dados los estadísticos de orden para una muestra aleatoria de tamaño n, $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$, se define el proceso estocástico

$$\xi_n(t) = \begin{cases} X_{[i]}, \ \frac{i-1}{n} \leq t \leq \frac{i}{n} \\ X_{[1]}, \ t = 0. \end{cases}$$

Su media dentro de cada clase de equivalencia se denota como $\overline{\xi}_n(t)$.

Por último, hay que especificar la función de pesos que se aplicará a la media del proceso anterior. Dicha función debe cumplir las siguientes propiedades:

i) $w(t) \geq 0$, $t \in [0,1]$;

ii) $\int_0^1 w(t)\, dt = 1$.

Además, de acuerdo con el objetivo de robustez para el *PL*-estadístico, es necesario que la función de pesos sea no creciente, ya que, debido a que la ordenación se hace de dentro hacia fuera, para valores de t próximos a uno no interesa asignar pesos mayores que a los puntos más internos.

Se define el *PL*-estadístico como $PL_n = \int_0^1 \overline{\xi}_n(t) w(t)\, dt$ o, equivalentemente, como $PL_n = \int_0^1 \xi_n(t) \overline{w}(t)\, dt$, donde $\overline{w}(t)$ es el peso medio dentro de cada clase de equivalencia, es decir, dada una clase de equivalencia con elementos $X_{[(i+1)/n]}, X_{[(i+2)/n]}, \ldots, X_{[(i+l)/n]}$ la función media de pesos será $\overline{w}(t) = \left(\frac{l}{n}\right)^{-1} \int_{i/n}^{(i+l)/n} w(s)\, ds$, para todo $t \in \left[\frac{i}{n}, \frac{i+l}{n}\right]$.

Un ejemplo ya mencionado de este tipo de estadísticos es el de las medias recortadas, en las que se asigna un peso nulo a las observaciones menos profundas. Si se desecha la proporción α de observaciones menos profundas, se tiene la media recortada por profundidad de $\alpha\,\%$. La función de pesos en este caso asigna el mismo valor para todas las

t menores que $1-\alpha$ y cero al resto: $w(t) = \frac{1}{1-\alpha} I_{[0,1-\alpha]}(t)$. En este caso particular se obtienen para los valores extremos $\alpha = 0$ y $\alpha = 1$ la media muestral y el punto más profundo o mediana respectivamente.

La versión poblacional se obtiene al promediar los puntos de la frontera de la región t-ésima definida en la sección (3.1) y que se denotó por $Q_F(t)$. La media de los puntos de esa curva será la media asociada al cuantil t-ésimo y se denotará por $\overline{Q}_F(t)$. Esta media es el equivalente de la cantidad muestral $\overline{\xi}_n(t)$. Dada la función de pesos, el valor del PL-estadístico será

$$PL_F = \int_0^1 \overline{Q}_n(t) w(t)\, dt.$$

El cálculo muestral de estadísticos de este tipo es trivial en caso de no haber empates entre puntos, ya que tras la ordenación de datos se promedian los $n(1-\alpha)$ puntos más profundos si dicha cantidad es entera y en caso contrario los $\lfloor n(1-\alpha) \rfloor + 1$ puntos, donde $\lfloor \cdot \rfloor$ denota la parte entera. Si hubiera empates y, por lo tanto, clases de equivalencia (más frecuentes para el método de la envolvente convexa por su definición) se asignará un peso $1/[n(1-\alpha)]$ a todas las observaciones anteriores a la clase de equivalencia frontera de $(1-\alpha)$ para la que se asigna el peso sobrante hasta la unidad y se reparte equitativamente sobre todos los elementos de la clase.

En Fraiman y Meloche (1999) se definen los L-estadísticos a partir de densidades estimadas, con el problema de elección del *ancho de banda*, y se comparan con los definidos a partir de la profundidad simplicial y otras medidas de centralidad.

1.5.4. Dispersión

Para el análisis de la variabilidad o dispersión de una variable aleatoria univariante es posible la obtención de medidas como la varianza, la desviación típica (y la desviación típica recortada, véase Bickel y Lehmann (1976)), la MEDA y el rango intercuartílico. Las dos últimas estiman la variabilid de forma más robusta que la primera debido al uso de estadísticos de orden para su cálculo. Si la variable aleatoria es multivariante es posible medir su dispersión a través de un escalar: por ejemplo, por medio de la varianza genera-

lizada que resume la información obtenida por la matriz de covarianzas. Sin embargo, si se analiza la variabilidad a través de un escalar se ignora mucha información importante que se extrae de la matriz de covarianzas. Esta información es la referente a las relaciones entre variables o, consecuentemente, la dirección sobre la que se distribuye la masa de probabilidad en el espacio y la anchura o grosor de la masa en torno a esta dirección. Por otro lado, si se consigue resumir la variabilidad en un escalar, la comparación de distribuciones atendiendo a la dispersión es más sencilla: es posible determinar si una distribución F está más o menos dispersa que otra G comparando tan sólo esa medida de variabilidad escalar. Este hecho no se da al comparar sus matrices de covarianzas que, a lo sumo, podrá hacerse componente a componente. Por ello, en un análisis de datos es necesaria la presencia de ambas formas de medir variabilidad, de ahí que sea necesaria una correcta definición y estimación de las medidas que harán posible el análisis.

En el estudio de la matriz de dispersión se puede emplear la estimación usual de la matriz de covarianzas, pero también es posible extender los estadísticos definidos en el apartado 1.5.3 con el objetivo conocido de eliminar el indeseable y posible efecto de observaciones extremas. Es necesaria, por tanto, la definición de la función de pesos que debe cumplir las propiedades comentadas anteriormente y la definición del proceso estocástico media. En este caso el proceso media estará formado por una matriz cuyos elementos serán media de procesos univariantes.

Se define el proceso $S_n(t)$ como la matriz obtenida del producto del vector distancia entre la observación (o clase de equivalencia) asociada a t con el centro υ_n que será elegido de acuerdo a la profundidad con que se haya realizado la ordenación

$$S_n(t) = \begin{cases} \left(X_{[i]} - \upsilon_n\right)\left(X_{[i]} - \upsilon_n\right)', & \text{para } \frac{i-1}{n} < t \leq \frac{i}{n} \\ O, & \text{para } t = 0, \end{cases}$$

donde la matriz O es una matriz de ceros.

A partir del proceso anterior se define su media por componentes dentro de cada clase de equivalencia y se denota por $\overline{S}_n(t)$. Por otro lado, entendiendo la integral de una matriz como la matriz de las integrales de cada componente, se expresa, a partir de $w(t)$, la siguiente estimación de matriz de dispersión:

$$S_n = \int_0^1 \overline{S}_n(t) w\,(t)\,dt = \int_0^1 S_n(t) \overline{w}\,(t)\,dt.$$

Como caso particular de este tipo de matrices de dispersión se obtiene para el peso unitario, $w\,(t) = 1$, para $0 \leq t \leq 1$, y $v_n = \overline{X}$ la matriz de dispersión clásica. Para funciones de ponderación del tipo $w\,(t) = \frac{1}{1-\alpha} I_{[0,1-\alpha]}\,(t)$ se obtendrá la matriz de dispersión muestral recortada de $\alpha\,\%$.

1.6. Otras aplicaciones

Las aplicaciones de las ideas de la profundidad estadística no se ciñen exclusivamente al análisis de datos. Por ejemplo, en Rousseeuw y Hubert (1999) se define una función de profundidad para rectas de regresión, se establecen cotas superiores e inferiores para el valor de la profundidad de la recta más profunda en la regresión simple y se propone un algoritmo para su cálculo. Haciendo uso de la recta más profunda se obtiene una estimación de la regresión más robusta que la mínimo cuadrática y L^1, ya que posee un punto de ruptura igual a 1/3. Además esa estimación es invariante ante transformaciones monótonas de las observaciones. En Mizera y Volauf (2002) se estudian las cotas del valor de la profundidad del hiperplano más profundo en regresión múltiple. En Van Aelst y Rousseeuw (2000) se demuestra que la dimensión no influye en el punto de ruptura. También en ese mismo trabajo se comprueba la consistencia de Fisher para dicho estimador. Existen algortimos exactos y aproximados para el cálculo del hiperplano más profundo en Van Aelst et al. (2002), donde se introduce también una metodología para los contrastes de hipótesis sobre los parámetros. Bai y He (1999) estudian el comportamiento asintótico del hiperplano de regresión por profundidad.

Las funciones de profundidad ha sido también aplicadas en la definición de contrastes de hipótesis, basándose, por ejemplo, en las posiciones de las observaciones ordenadas y no en su valor, en Liu y Singh (1993) se desarrollan contrastes de hipótesis sobre la media y la dispersión aplicándolos sobre índices de calidad, con el fin de determinar si una muestra obtenida de un proceso de fabricación proviene o no de una supuesta distribución. También, en Liu y Singh (1997), donde de se aplica el bootstrap para verificar

la “verosimilitud” de cierto vector de parámetros, comparando el rango que éste ocupa dentro de la muestra bootstrap de puntos más profundos. Más aplicaciones en contrastes pueden encontrarse en Liu y Singh (1992), Liu et al. (1999) y Hettmansperger et al. (1994). Aplicaciones del bootstrap sobre los puntos más profundos pueden encontrarse en Yeh y Singh (1997), en el que se construyen regiones de confianza balanceadas para el vector de parámetros. Y aplicaciones en análisis de conglomerados y clasificación en Jornsten (2004), Jornsten et al. (2002) y López-Pintado y Romo (2007).

Capítulo 2

Similaridades basadas en profundidad

Resumen

En este capítulo se proponen funciones para medir la proximidad o similaridad entre puntos en un sentido estadístico, es decir, del mismo modo en que las funciones de profundidad miden la centralidad de los puntos: teniendo en cuenta la forma del conjunto de puntos o de la función de distribución generadora. Aunque, en muchos de los ejemplos que se introducen a continuación, las funciones pueden extenderse de forma trivial a casos en que interese comparar más de dos puntos simultáneamente, este capítulo se centra en el estudio de las proximidades entre pares de puntos. En la primera sección se presenta la generalización de algunas de las funciones de profundidad propuestas en la literatura. En la segunda se muestran, para cada una de las similaridades que se definen, algunos gráficos que ilustran el funcionamiento sobre dos conjuntos de datos simulados, uno generado de una distribución simétrica y otro de una asimétrica. La generalización de las funciones de profundidad depende de su forma funcional, de ahí que las propiedades que se estudian en la tercera sección estén muy determinadas por las de la función original. Estas propiedades se dividen en dos grupos. El primero recoge el estudio de una extensión de las propiedades deseables de las funciones de profundidad. El segundo grupo está formado por las propiedades asintóticas de la versión muestral y de continuidad para

distribuciones continuas. La última sección contiene la aplicación de las similaridades en el análisis de conglomerados jerárquicos. Se muestran los grupos que se obtienen con cada una de éstas y se comparan con los que se obtienen empleando la distancia euclídea, obteniéndose que todas las similaridades excepto una, mejoran los resultados obtenidos con ésta.

2.1. Medidas de proximidad

Las medidas de proximidad se emplean para cuantificar la cercanía o lejanía de puntos, observaciones, individuos o variables en el espacio. Estas medidas pueden clasificarse en dos grupos: el de similaridades s compuesto por funciones que miden la cercanía de dos puntos (lo parecidos que son dos objetos) y el de disimilaridades δ con funciones que miden la lejanía de dos puntos (lo que se diferencian dos objetos).

Para cualquier par de puntos x e y en $\mathbb{R}^d$, se tiene que, tanto las similaridades como las disimilaridades, toman valores no negativos. Además, la disimilaridad entre un punto x y él mismo es igual a cero. No sucede lo mismo para las similaridades ya que, aunque generalmente están escaladas y el valor de la similaridad entre un punto y él mismo es igual a uno, puede tomar valores distintos. A pesar de esto, sí se tiene que $\max_{y\in\mathbb{R}^d} s_{x,y} = s_{x,x}$. Otra propiedad que habitualmente se exige a las proximidades es la de simetría.

A continuación se enumeran algunos ejemplos de medidas de proximidad en $\mathbb{R}^d$:

- **Distancia euclídea**: Dados los vectores $x = (x^{1)}, x^{2)}, \ldots, x^{d)})$ e $y = (y^{1)}, y^{2)}, \ldots, y^{d)})$, se define la distancia euclídea como

$$d(x, y) = \|x - y\|,$$

siendo $\|x\| = \sqrt{\sum_{i=1}^{d} \left(x^{i)}\right)^2}$.

- **Distancia euclídea ponderada**: Se define como la distancia euclídea, pero ponderando por el vector $\omega = (\omega^{1)}, \omega^{2)}, \ldots, \omega^{d)})$, $\omega^{i)} \geq 0$,

$$d_\omega(x, y) = \sqrt{\sum_{i=1}^{d} \omega^{i)} \left(x^{i)} - y^{i)}\right)^2}.$$

- **Distancia de Mahalanobis**: Dados los vectores d-dimensionales x e y y una matriz de varianzas-covarianzas no singular Σ, se define como

$$d_{Mah}(x, y) = \sqrt{(x - y)' \, \Sigma^{-1} \, (x - y)}.$$

- **Separación angular**: Mide el coseno del ángulo que separa a los dos vectores. Se define como

$$d_{ang}(x,y) = \frac{\sum_{i=1}^{d} x^{i)} y^{i)}}{\|x\| \, \|y\|}.$$

- **Distancia de cuerda**: Proyecta los puntos sobre el círculo unidad y calcula la distancia euclídea entre los puntos proyectados. Está definida como

$$d_{cuerda}(x,y) = \sqrt{\sum_{i=1}^{d} \left(\frac{x^{i)}}{\|x\|} - \frac{y^{i)}}{\|y\|} \right)^2}.$$

- **Correlación de Pearson**: Mide la asociación lineal existente entre las componentes de los dos vectores. Se define como

$$\rho(x,y) = \frac{\sum_{i=1}^{d} (x^{i)} - \overline{x})(y^{i)} - \overline{y})}{\sqrt{\sum_{i=1}^{d} (x^{i)} - \overline{x})^2 \sum_{i=1}^{d} (y^{i)} - \overline{y})^2}}.$$

De todas estas funciones tan sólo la última se encuadra dentro del grupo de similaridades. Es sobre este grupo sobre el que se trabaja en este capítulo para la definición de nuevas similaridades. Las propiedades a exigir a estas nuevas funciones desde el punto de vista de similaridad son las siguientes:

(i) No negatividad: $s_{x,y} \geq 0$

(ii) Maximalidad sobre los puntos: $\max_{y \in \mathbb{R}^d} s_{x,y} = s_{x,x}$ y $\max_{x \in \mathbb{R}^d} s_{x,y} = s_{y,y}$

(iii) Simetría: $s_{x,y} = s_{y,x}$

2.2. Funciones de similaridad

La interpretación de las funciones de profundidad admite varias versiones equivalentes. Por un lado, tal y como se comentó en el capítulo introductorio, pueden entenderse como medidas del grado de centralidad de puntos con respecto a una función de distribución. Por otro lado, está la que motiva este capítulo, que la considera como una medida de similaridad entre puntos y centro. La Figura 2.1 contiene las curvas de nivel de la profundidad

simplicial aplicada a una muestra simulada de una normal bivariante. El punto marcado en el centro de la Figura se corresponde con el punto más profundo de la muestra. La disimilaridad entre éste y el otro punto del dibujo puede entenderse como el número de contornos que hay entre ambos puntos. Puntos más alejados del centro que éste en esa dirección tendrán un mayor número de contornos entre ambos. También se observa que la proximidad en este sentido no es igual en todas las direcciones ya que, sobre otras direcciones, es posible encontrar puntos que estén a la misma distancia euclídea y que sean más (o menos) centrales que el punto marcado. Por lo tanto, se tiene que la profundidad se adapta a la forma de la nube de puntos. Así pues, la idea sería entender la proximidad como la masa de probabilidad que hay entre ambos puntos (poca masa, puntos próximos).

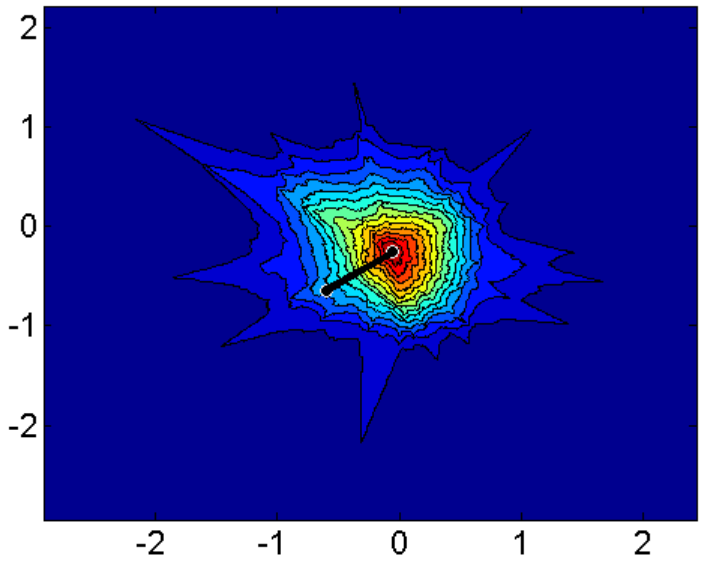


Figura 2.1: *Profundidad simplicial vista como medida de similaridad.*

Las funciones de similaridad basadas en profundidad que se introducen en este capítulo son generalizaciones de funciones de profundidad de los tipos A, B y C según la clasificación propuesta en Zuo y Serfling (2000a). Es importante tener en cuenta dicha clasificación ya que, tanto la generalización para la obtención de similaridades como las posibilidades de aplicación a más de dos puntos, dependen totalmente del tipo de función

de profundidad elegido. Del tipo A se estudian aquí las profundidades de Mahalanobis y por proyecciones; del tipo B la profundidad de Oja; y del tipo C las profundidades simplicial, por bandas y por bandas modificada.

Notación 2.1 *Sean x e y dos puntos en $\mathbb{R}^d$ y F una función de distribución d-dimensional. La proximidad o similaridad basada en profundidad (S) entre x e y con respecto a la función de distribución F se denota como $S(x, y; F)$ o $S_F(x, y)$.*

Notación 2.2 *Sean x e y dos puntos en $\mathbb{R}^d$ y F_n la función de distribución empírica de una muestra aleatoria simple de tamaño n de la función de distribución F. La proximidad o similaridad basada en profundidad muestral entre x e y con respecto a la función de distribución F_n se denota como $S_n(x, y)$.*

2.2.1. Similaridad de Mahalanobis

Se construye empleando la distancia de Mahalanobis. La profundidad emplea la distancia entre punto y esperanza de la distribución. En la similaridad se sustituye la esperanza μ por el otro punto que se desea comparar, es decir, considerando la distancia de Mahalanobis entre ambos puntos. Esta función mide proximidades de manera adecuada siempre que la forma de la nube de puntos (o distribución) sea elíptica. Además necesita la existencia de los dos primeros momentos de la distribución.

Definición 2.1 *Sea el vector aleatorio d-dimensional X con función de distribución F y con matriz de varianzas y covarianzas $\Sigma_F = E\left[(X - E[X])(X - E[X])'\right]$. Dados dos puntos x e y en $\mathbb{R}^d$, se define la similaridad de Mahalanobis con respecto a F como*

$$SM(x, y; F) = \left[1 + (x - y)' \Sigma_F^{-1} (x - y)\right]^{-1}.$$

Si $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ es una muestra aleatoria de F, la versión muestral se obtiene sustituyendo la matriz de varianzas y covarianzas por una estimación suya. La robustez de la versión muestral de esta similaridad depende de la del estimador de la matriz de varianzas y covarianzas empleado.

2.2.2. Similaridad por proyecciones

Consiste en una reducción de la dimensión de las observaciones a un espacio unidimensional empleando proyecciones sobre vectores de norma dos igual a uno. Sobre los puntos proyectados se considera la distancia en valor absoluto y se estandariza por la variabilidad de la distribución proyectada. Ésta se estima mediante la MEDA o mediana de los valores absolutos de las desviaciones respecto de la mediana. Tomando el mayor valor posible de esa distancia estandarizada se obtiene la medida de separación entre ambos puntos o medida de atipicidad con respecto a F,

$$A(x, y; F) = \sup_{\|u\|=1} \frac{|u'x - u'y|}{MEDA(u'X)}.$$

Esta función toma valores reales no negativos, por lo que es necesario aplicar una transformación, como en el caso anterior, para obtener una medida de similaridad análoga a la función de profundidad de la que es extensión.

Definición 2.2 *Sea el vector aleatorio d-dimensional X con función de distribución F. Dados dos puntos x e y en $\mathbb{R}^d$, se define la similaridad por proyecciones con respecto a F como*

$$SP(x, y; F) = \left[1 + \sup_{\|u\|=1} \frac{|u'x - u'y|}{MEDA(u'X)}\right]^{-1}.$$

Si $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ es una muestra aleatoria de F, las similaridades muestrales se obtienen sustituyendo la $MEDA(u'X)$ por una estimación calculada a partir de la muestra proyectada, $u'x_1, u'x_2, \ldots, u'x_n$. La versión muestral es más robusta que la correspondiente a la similaridad de Mahalanobis. En cuanto a su forma, fijado uno de los dos puntos, por ejemplo, x, la similaridad verifica para todo $z \in \mathbb{R}^d$ y para todo $\alpha \geq 0$ que $SP(x + \alpha z, x; F) = SP(x - \alpha z, x; F)$.

2.2.3. Similaridad de Oja

La similaridad de Oja se construye a partir de volúmenes de símplices de tamaño $d+1$, donde d es la dimensión del espacio que contiene a las observaciones. En la función de profundidad, uno de los vértices de los símplices es siempre el punto x. Para construir

la similaridad se incluye como vértice de todos los símplices además de a éste, al punto y. El resto de los vértices de los símplices se completan de forma aleatoria con $d-1$ copias de la distribución F. Como puede observarse con los ejemplos de la Figura 2.2, cuando los puntos x e y (círculos de color rojo) están próximos (gráfico de la izquierda), el volumen medio de los símplices para los que estos puntos son vértices es próximo a cero, a diferencia de cuando los puntos están alejados (gráfico de la derecha).

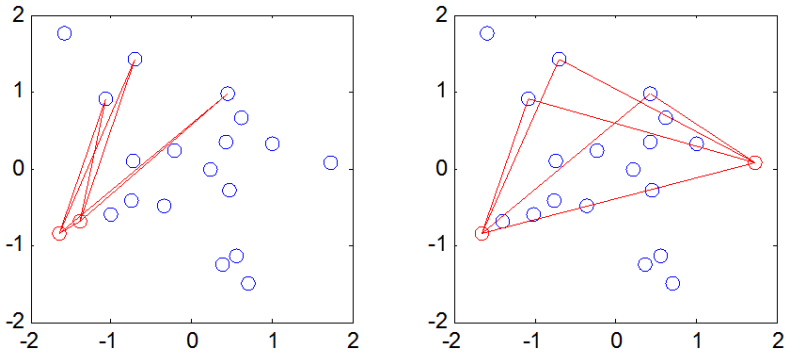


Figura 2.2: *Ejemplos de símplices generados aleatoriamente con vértices próximos y* distantes.

De nuevo, dado que el volumen puede tomar cualquier valor real no negativo, se realiza la misma transformación que en las similaridades anteriores.

Definición 2.3 *Sea la función de distribución d-dimensional F. Dados dos puntos x e y en $\mathbb{R}^d$, se define la similaridad de Oja entre x e y con respecto a F como*

$$SO(x, y; F) = \left[1 + E_F\left(Vol(S[x, y, X_1, X_2, \ldots, X_{d-1}])\right)\right]^{-1},$$

donde $X_1, X_2, \ldots, X_{d-1}$ son variables aleatorias independientes con función de distribución F.

Esta similaridad será igual a uno para todas las distribuciones discretas que tengan menos de $d-1$ puntos con probabilidad no nula, ya que el volumen de todos los símplices aleatorios es nulo.

Si $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ es una muestra aleatoria de F, la versión muestral de esta similaridad se obtiene promediando el volumen de todos los posibles símplices formados

por los puntos x e y, y por $d-1$ puntos de la muestra

$$SO_n(x,y) = \left[1 + \binom{n}{d-1}^{-1} \sum_{1 \le i_1 < \ldots < i_{d-1} \le n} Vol(S[x, y, x_{i_1}, x_{i_2}, \ldots, x_{i_{d-1}}])\right]^{-1}.$$

Esta similaridad permite de forma trivial su extensión para la comparación de más de dos puntos simultáneamente. Partiendo de la profundidad de Oja, la similaridad ha sido construida eliminando un vértice aleatorio e incluyendo el nuevo punto y. Siguiendo este esquema se podría llegar a medir la proximidad de un máximo de $d+1$ puntos. En este caso extremo, la similaridad no dependería de la distribución ya que no habría ningún término aleatorio dentro de los símplices.

2.2.4. Similaridad simplicial

Esta similaridad también está basada en símplices con $d+1$ vértices. La profundidad simplicial mide la probabilidad de que un símplice aleatorio contenga al punto para el que se calcula la profundidad. La extensión que se propone aquí para dos puntos consiste en exigir la pertenencia de ambos puntos al símplice. En la Figura 2.3 se presentan dos ejemplos que muestran que la propuesta es una forma adecuada para medir proximidades. Se puede observar que, cuando los puntos x e y (círculos en rojo) están próximos, de los cinco triángulos generados aleatoriamente que aparecen, dos de ellos (triángulos en color rojo) contienen a ambos puntos a la vez, mientras que si los puntos están alejados, ningún triángulo de estos cinco los contiene simultáneamente.

Definición 2.4 *Sea la función de distribución d-dimensional F. Dados dos puntos x e y en $\mathbb{R}^d$, se define la similaridad simplicial entre x e y con respecto a F como*

$$SS(x, y; F) = Pr(x, y \in S[X_1, X_2, \ldots, X_{d+1}]),$$

donde $X_1, X_2, \ldots, X_{d+1}$ son variables aleatorias independientes e idénticamente distribuidas según F.

Observación 2.1 *Esta similaridad puede escribirse también como*

$$SS(x, y; F) = E_F\{h(x, y; X_1, X_2, \ldots, X_{d+1})\},$$

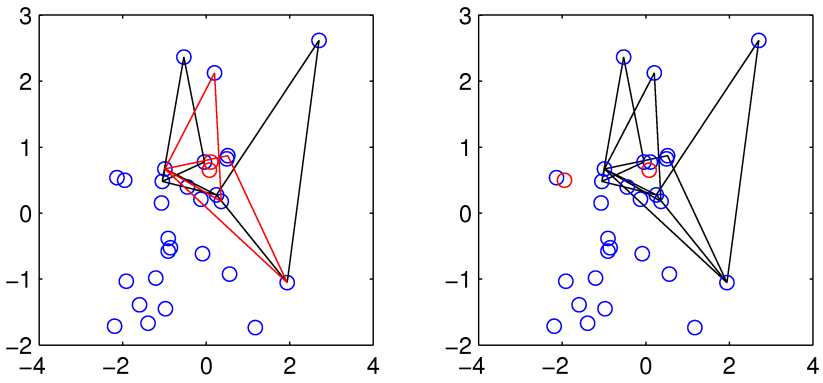


Figura 2.3: *Ejemplos de símplices generados aleatoriamente para puntos próximos y distantes.*

donde $h\left(x, y; X_1, X_2, \ldots, X_{d+1}\right)$ *es la función indicadora de inclusión de* x *e* y *en el símplice,* $I\left(x, y \in S\left[X_1, X_2, \ldots, X_{d+1}\right]\right)$.

Un concepto estadístico empleado tanto en las funciones de profundidad, como en las similaridades cuya forma funcional es análoga a la simplicial, es la noción de U-estadístico (Hoeffding (1948)) que se introduce a continuación.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria en $\mathbb{R}^d$ con distribución F. Dada una función m-dimensional $h\left(x_1, \ldots, x_m\right)$ denominada núcleo, el parámetro $\theta\left(F\right) = E_F\left[h\left(X_1, \ldots, X_m\right)\right]$, se estima a través de su correspondiente U-estadístico que, a partir de una muestra $X_1, X_2, \ldots, X_n$, con $m \leq n$, se obtiene como

$$U_n = U\left(X_1, X_2, \ldots, X_n\right) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 \leq \ldots \leq i_m \leq n} h\left(X_{i_1}, \ldots, X_{i_m}\right).$$

Si $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ es una muestra aleatoria de F, la versión muestral de esta similaridad es un *U-estadístico* con función núcleo $h\left(x, y; x_1, x_2, \ldots, x_{d+1}\right)$. El cálculo de la esperanza de esta función se realiza promediando la pertenencia sobre todos los conjuntos de $d+1$ elementos de las n observaciones, es decir,

$$SS_n(x, y) = \binom{n}{d+1}^{-1} \sum_{1 \leq i_1 < \ldots < i_{d+1} \leq n} I(x, y \in S[x_{i_1}, x_{i_2}, \ldots, x_{i_{d+1}}]).$$

Esta similaridad puede también ser ampliada de manera trivial para la comparación de más de dos puntos. No presenta la limitación técnica comentada para la similaridad de Oja en cuanto al número de puntos a comparar. Sin embargo, sí presenta una limitación práctica: si el número de puntos comparados representa un porcentaje alto de la muestra apenas habrá símplices que los contengan a todos (especialmente si alguno de los puntos está algo alejado del resto) y, por lo tanto, la mayoría de las comparaciones tendrán valores de similaridades muy bajos o nulos.

2.2.5. Similaridad por bandas

En conjuntos de observaciones de alta dimensión, aquellas similaridades que tienen factores combinatorios dependientes de dicha dimensión requieren un elevado tiempo de cómputo. No ocurre lo mismo para la similaridad por bandas, obtenida a partir de una profundidad cuyos requerimientos en tiempo de cálculo son notablemente inferiores (véase López-Pintado y Romo (2009)). Para ilustrar la idea se emplea el sistema de coordenadas paralelas, en el que el eje X contiene el número de coordenada y el eje Y el valor de la variable para cada coordenada. La Figura 2.4 muestra la representación en coordenadas paralelas de los puntos $x = (1, 2.5, 2, 3.5, 2.5)$ (en rojo) e $y = (1.5, 4, 4, 4.5, 3)$ (en azul).

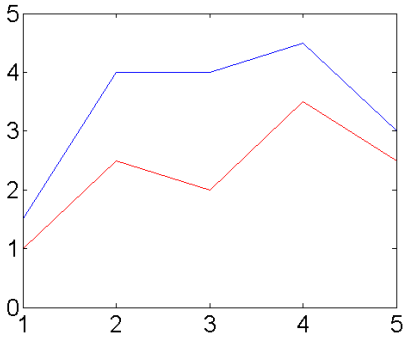


Figura 2.4: *Representación en coordenadas paralelas de los puntos x e y.*

La región comprendida entre los grafos de esos dos puntos de la Figura se denomina banda formada por x e y. Las bandas pueden estar formadas por dos o más puntos. La Figura 2.5 muestra un ejemplo de una banda formada por cuatro observaciones de

dimensión diez.

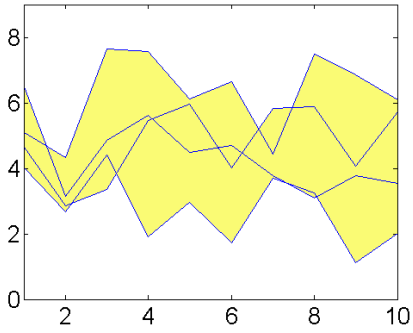


Figura 2.5: *Banda formada por cuatro puntos.*

Más formalmente se define la banda determinada por los puntos $x_1, x_2, \ldots, x_b \in \mathbb{R}^d$ como

$$B\left(x_1, x_2, \ldots, x_b\right) = \left\{ y \in \mathbb{R}^d : \forall k \in \{1, 2, \ldots, d\}, \min_{i \in \{1,\ldots,b\}} x_i^{k)} \leq y^{k)} \leq \max_{i \in \{1,\ldots,b\}} x_i^{k)} \right\},$$

donde $x_i^{k)}$ es la coordenada k-ésima del punto x_i e $y^{k)}$ la coordenada k-ésima del punto y.

La similaridad por bandas entre dos puntos se define, al igual que la similaridad simplicial, mediante la pertenencia de manera simultánea de ambos puntos a regiones aleatorias (bandas). La Figura 2.6 contiene dos ejemplos que ilustran el funcionamiento de esta similaridad. En ambos ejemplos se encuentra representada en coordenadas paralelas una muestra de quince puntos de dimensión diez. Las curvas en rojo son aquellas para las que se ilustra la similaridad. Las curvas de color verde tienen para todas sus coordenadas valores mayores que las coordenadas de los puntos a comparar, mientras que las moradas tienen todas sus coordenadas menores. Cualquier combinación de curvas que contenga al menos una verde y otra morada dará lugar a una banda que contiene a ambos puntos. Se puede observar que el número de curvas candidatas a bandas que incluyan a ambos puntos es mayor si los puntos están próximos (gráfico de la izquierda).

Notación 2.3 *Para bandas formadas por un número b de puntos de $\mathbb{R}^d$, dados dos puntos x e y en $\mathbb{R}^d$ y una función de distribución d-dimensional F, se denota por $SB^{b)}\left(x, y; F\right)$ a*

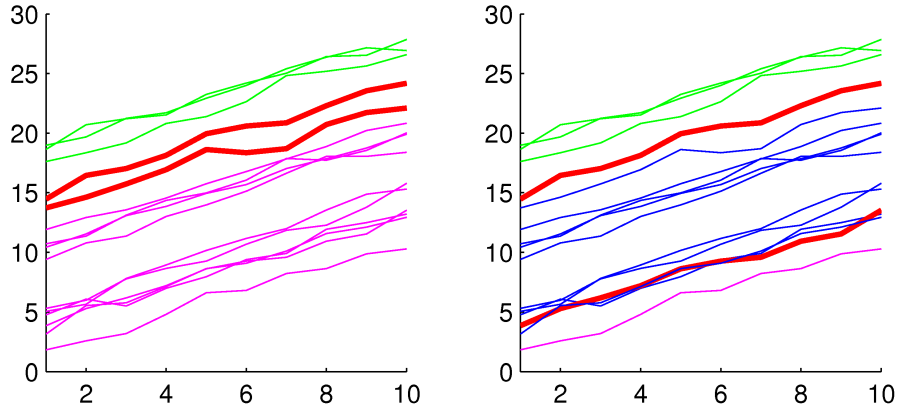


Figura 2.6: *Ejemplos de candidatos a bandas para puntos próximos y distantes.*

la probabilidad de que bandas aleatorias formadas por b puntos con función de distribución F contengan a los puntos x e y; más formalmente,

$$\begin{aligned} SB^{b)}(x,y;F) &= Pr\left(x,y \in B\left(X_1,X_2,\ldots,X_b\right)\right) \\ &= E\left[\prod_{k=1}^{d} I\left\{\min_{i\in\{1,2,\ldots,b\}} X_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{1,2,\ldots,b\}} X_i^{k)}\right\}\right], \end{aligned}$$

donde $x^{k)}$ es la coordenada k-ésima de x, $X_i^{k)}$ la coordenada k de la variable aleatoria X_i y X_i, con $i = 1,2,\ldots,b$, son variables aleatorias independientes e idénticamente distribuidas según F.

Definición 2.5 *Dada la función de distribución d-dimensional F, el número máximo B de puntos para las bandas y dos puntos x e y en $\mathbb{R}^d$, se define la similaridad por bandas entre x e y con respecto a F como*

$$SB\left(x,y;F,B\right) = \sum_{b=2}^{B} SB^{b)}\left(x,y;F\right), \ B \geq 2.$$

Si $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ es una muestra aleatoria de F, la versión muestral de la similaridad se obtiene sumando la estimación de $SB^{b)}\left(x,y;F\right)$ (denotada por $SB_n^{b)}\left(x,y;F\right)$) para bandas formadas por b puntos donde $b = 2,\ldots,B$,

$$SB_n\left(x,y;B\right) = \sum_{b=2}^{B} SB_n^{b)}\left(x,y\right), \ B \geq 2,$$

donde la cantidad $SB_n^{b)}(x,y)$ es resultado de promediar la pertenencia a una banda sobre todas las generadas por b puntos que se pueden formar con observaciones de la muestra

$$SB_n^{b)}(x,y) = \binom{n}{b}^{-1} \sum_{1\leq i_1<\ldots<i_b\leq n} \prod_{k=1}^{d} I\left\{ \min_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)} \right\}.$$

Cuando la dimensión es elevada en relación al número de observaciones de la muestra, o cuando las componentes de los vectores aleatorios no tienen una alta dependencia entre ellas, las bandas que esta similaridad emplea son muy restrictivas, lo que puede dar lugar a problemas en la estimación de las proximidades.

2.2.6. Similaridad por bandas modificada

La similaridad por bandas modificada, basada en la profundidad por bandas modificada (véase López-Pintado y Romo (2009)), es una similaridad más flexible que la anterior. Ésta no mide el porcentaje de bandas que contienen a los puntos x e y, sino la esperanza del porcentaje de coordenadas para las que ambos puntos están dentro de las bandas aleatorias, es decir, dados los puntos $x_1, x_2, \ldots, x_b$ que conforman una banda y los puntos x e y, se mide

$$\#\left(C\left(x,y;x_1,x_2,\ldots,x_b\right)\right),$$

donde

$$C\left(x,y;x_1,x_2,\ldots,x_b\right) = \left\{ k\in\{1,2,\ldots,d\} : \min_{i\in\{1,2,\ldots,b\}} x_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{1,2,\ldots,b\}} x_i^{k)} \right\}$$

representa las coordenadas para las que los puntos x e y están dentro de la banda formada por los puntos $x_1,\ldots,x_b$ y $\#(C)$ es el cardinal del conjunto C (el número de coordenadas para las que la banda contiene a ambos puntos).

Notación 2.4 *Para bandas formadas por un número b de puntos de $\mathbb{R}^d$, dados dos puntos x e y en $\mathbb{R}^d$ y una función de distribución d-dimensional F, se denota por $SBM^{b)}(x,y;F)$ el porcentaje medio de coordenadas para las que los puntos x e y están dentro de bandas*

aleatorias formadas por b puntos, es decir,

$$\begin{aligned} SBM^{b)}(x,y;F) &= \frac{1}{d}E\left[\#\left(C\left(x,y;X_1,X_2,\ldots,X_b\right)\right)\right] \\ &= E\left[\frac{1}{d}\sum_{k=1}^{d} I\left\{\min_{i\in\{1,2,\ldots,b\}} X_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{1,2,\ldots,b\}} X_i^{k)}\right\}\right] \\ &= \frac{1}{d}\sum_{k=1}^{d} E\left[I\left\{\min_{i\in\{1,2,\ldots,b\}} X_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{1,2,\ldots,b\}} X_i^{k)}\right\}\right] \\ &= \frac{1}{d}\sum_{k=1}^{d} Pr\left[\min_{i\in\{1,2,\ldots,b\}} X_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{1,2,\ldots,b\}} X_i^{k)}\right], \end{aligned}$$

donde $x^{k)}$, $X_i^{k)}$ e $y^{k)}$ son, respectivamente, las coordenadas k-ésimas de x, X_i e y, y X_i, con $i = 1, 2, \ldots, b$, son variables aleatorias independientes e idénticamente distribuidas según F.

Definición 2.6 *Dada la función de distribución d-dimensional F, el número máximo B de puntos para las bandas y dos puntos x e y en $\mathbb{R}^d$, se define la similaridad por bandas modificada entre x e y con respecto a F como*

$$SBM\left(x,y;F,B\right) = \sum_{b=2}^{B} SBM^{b)}\left(x,y;F\right),\ B \geq 2.$$

Si $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ es una muestra aleatoria de F, su versión muestral se obtiene sumando las probabilidades de pertenencia estimadas para bandas generadas con un máximo de B puntos de la muestra, es decir,

$$SBM_n\left(x,y;B\right) = \sum_{b=2}^{B} SBM_n^{b)}\left(x,y\right),\ B \geq 2,$$

donde la probabilidad de pertenencia se estima, para cada b, promediando sobre todas las posibles bandas de b puntos

$$SBM_n^{b)}(x,y) = \binom{n}{b}^{-1} \sum_{1\leq i1<\ldots<i_b\leq n} d^{-1}\sum_{k=1}^{d} I\left\{\min_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)}\right\}.$$

Observación 2.2 *Dada una función de distribución d-dimensional F y dos puntos x e $y \in \mathbb{R}^d$, si se denota por $SBM^{b,k)}\left(x^{k)}, y^{k)}; F\right)$ la expresión*

$$E\left[I\left\{\min_{i\in\{i_1,i_2,\ldots,i_b\}} X_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{i_1,i_2,\ldots,i_b\}} X_i^{k)}\right\}\right],$$

donde $X_i^{k)}$ *es la* k*-ésima coordenada de la variable aleatoria* X_i *con distribución* F *y* $x^{k)}$ *la* k*-ésima coordenada de* x*, entonces la similaridad por bandas modificada puede reescribirse como*

$$SBM\left(x, y; B, F\right) = d^{-1} \sum_{b=2}^{B} \sum_{k=1}^{d} SBM^{b,k)}\left(x^{k)}, y^{k)}; F\right).$$

Observación 2.3 *Dada una muestra aleatoria* $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ *de la función de distribución* F *y dos puntos* x *e* $y \in \mathbb{R}^d$*, si se denota por* $SBM_n^{b,k)}\left(x^{k)}, y^{k)}\right)$ *la expresión*

$$\binom{n}{b}^{-1} \sum_{1 \leq i_1 < \ldots < i_b \leq n} I\left\{ \min_{i \in \{i_1, i_2, \ldots, i_b\}} x_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i \in \{i_1, i_2, \ldots, i_b\}} x_i^{k)} \right\},$$

donde $x_i^{k)}$ *es la* k*-ésima coordenada del punto* x_i *y* $x^{k)}$ *la* k*-ésima coordenada de* x*, entonces la similaridad por bandas modificada muestral puede reescribirse como*

$$SBM_n\left(x, y; B\right) = d^{-1} \sum_{b=2}^{B} \sum_{k=1}^{d} SBM_n^{b,k)}\left(x^{k)}, y^{k)}\right).$$

Un algoritmo *naive* para el cálculo de esta similaridad consiste en construir todas las posibles bandas e ir realizando, para cada par de puntos y cada banda, los cálculos necesarios. A pesar de que el número de posibles bandas, fijado un valor de B, sólo crece con el tamaño muestral, es decir, no depende de la dimensión de las observaciones, el tiempo de cómputo puede ser elevado si se realizan numerosas repeticiones. El siguiente resultado se ha obtenido con el fin de mejorar el rendimiento del cálculo de las similaridad muestral.

Teorema 2.1 *Sea la muestra* $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ *y las siguientes matrices*

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{pmatrix} \; y \; L = \begin{pmatrix} l_{1,1} & \cdots & l_{1,d} \\ \vdots & & \vdots \\ l_{n,1} & \cdots & l_{n,d} \end{pmatrix},$$

donde $l_{i,k} = \#\left\{x_{s,k} : x_{s,k} < x_{i,k}, s = 1, 2, \ldots, n\right\}$ *es el número de elementos de la columna* k *de la matriz* X *con valores menores que la coordenada* k *de la observación* i*. Entonces, la similaridad por bandas modificada entre dos puntos de la muestra* x_i *y* x_j *con respecto a* F_n *es igual a*

$$SBM_n\left(x_i, x_j; B\right) = d^{-1} \sum_{k=1}^{d} \sum_{b=2}^{B} SBM_n^{b,k)}\left(x_{i,k}, x_{j,k}\right)$$

donde $x_{i,k}$ es la coordenada k-ésima de la observación i y la función $SBM_n^{b,k}(x_{i,k}, x_{j,k})$se calcula, si $x_{i,k} \neq x_{j,k}$, según la expresión

$$\begin{aligned} SBM_n^{b,k)}(x_{i,k}, x_{j,k}) &= \binom{n}{b}^{-1} \left[l_{m_{i,j,k}} \left(n - l_{M_{i,j,k}} - \eta_{M_{i,j,k}} \right) \binom{n-2}{b-s} \right. \\ &+ \eta_{m_{i,j,k}} \left(n - l_{M_{i,j,k}} - \eta_{M_{i,j,k}} \right) \binom{n - l_{m_{i,j,k}} - 2}{b-2} \\ &+ \eta_{M_{i,j,k}} l_{m_{i,j,k}} \binom{l_{M_{i,j,k}} + \eta_{M_{i,j,k}} - 2}{B-2} \\ &\left. + \eta_{m_{i,j,k}} \eta_{M_{i,j,k}} \binom{l_{M_{i,j,k}} - l_{m_{i,j,k}} + \eta_{M_{i,j,k}} - 2}{B-2} \right], \end{aligned}$$

donde $\eta_{m_{i,j,k}}$ es la multiplicidad del $\min(x_{i,k}, x_{j,k})$ dentro de la columna k, $\eta_{M_{i,j,k}}$ la del $\max(x_{i,k}, x_{j,k})$, $l_{m_{i,j,k}} = \min(l_{i,k}, l_{j,k})$, $l_{M_{i,j,k}} = \max(l_{i,k}, l_{j,k})$ y $\binom{a}{b} = 0$ cuando $a < b$. Si se tiene que $x_{i,k} = x_{j,k}$, entonces la función $SBM_n^{b,k)}(x_{i,k}, x_{j,k})$ se calcula según la expresión

$$\begin{aligned} SBM_n^{b,k)}(x_{i,k}, x_{j,k}) &= \binom{n}{b}^{-1} \left[l_{i,k} \left(n - l_{i,k} - \eta_{i,k} \right) \binom{n-2}{b-s} \right. \\ &+ \eta_{i,j} \left(n - l_{i,k} - \eta_{i,k} \right) \binom{n - l_{i,k} - 2}{b-2} \\ &+ \eta_{i,k} l_{i,k} \binom{l_{i,k} + \eta_{i,k} - 2}{B-2} \\ &\left. + \binom{\eta_{i,k}}{2} \binom{l_{i,k} - l_{i,k} + \eta_{i,k} - 2}{B-2} \right], \end{aligned}$$

donde $\eta_{i,k}$ es la multiplicidad de $x_{i,k}$ dentro de la columna k, $l_{i,k} = \#\{x_{s,k} : x_{s,k} < x_{i,k}, s = 1, 2, \ldots, n\}$ y se supone que $\binom{a}{b} = 0$ cuando $a < b$.

Demostración. Denotando $m_{i,j,k} = \min(x_{i,k}, x_{j,k})$ y $M_{i,j,k} = \max(x_{i,k}, x_{j,k})$ se desarrolla la expresión del número de bandas que contienen a las coordenadas

$$\begin{aligned} &\sum_{1 \leq i_1 < \ldots < i_b \leq n} I \left\{ \min_{t \in \{i_1, i_2, \ldots, i_b\}} x_{t,k} \leq x_{i,k}, x_{j,k} \leq \max_{t \in \{i_1, i_2, \ldots, i_b\}} x_{t,k} \right\} \\ &= \sum_{1 \leq i_1 < \ldots < i_b \leq n} I \left\{ \min_{t \in \{i_1, i_2, \ldots, i_b\}} x_{t,k} \leq x_{i,k}, x_{j,k} \right\} I \left\{ \max_{t \in \{i_1, i_2, \ldots, i_b\}} x_{t,k} \geq x_{i,k}, x_{j,k} \right\} \\ &= \sum_{1 \leq i_1 < \ldots < i_b \leq n} I \left\{ \min_{t \in \{i_1, i_2, \ldots, i_b\}} x_{t,k} \leq \min(x_{i,k}, x_{j,k}) \right\} I \left\{ \max_{t \in \{i_1, i_2, \ldots, i_b\}} x_{t,k} \geq \max(x_{i,k}, x_{j,k}) \right\} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{1\le i_1<\ldots<i_b\le n} I\left\{\min_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} \le m_{i,j,k}\right\} I\left\{\max_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} \ge M_{i,j,k}\right\} \\
&= \sum_{1\le i_1<\ldots<i_b\le n} \left(I\left\{\min_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} < m_{i,j,k}\right\} + I\left\{\min_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} = m_{i,j,k}\right\}\right) \times \\
&\quad \left(I\left\{\max_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} > M_{i,j,k}\right\} + I\left\{\max_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} = M_{i,j,k}\right\}\right) \\
&= \sum_{1\le i_1<\ldots<i_b\le n} I\left\{\min_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} < m_{i,j,k}\right\} I\left\{\max_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} > M_{i,j,k}\right\} \\
&\quad + \sum_{1\le i_1<\ldots<i_b\le n} I\left\{\min_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} < m_{i,j,k}\right\} I\left\{\max_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} = M_{i,j,k}\right\} \\
&\quad + \sum_{1\le i_1<\ldots<i_b\le n} I\left\{\min_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} = m_{i,j,k}\right\} I\left\{\max_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} > M_{i,j,k}\right\} \\
&\quad + \sum_{1\le i_1<\ldots<i_b\le n} I\left\{\min_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} = m_{i,j,k}\right\} I\left\{\max_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} = M_{i,j,k}\right\}.
\end{aligned}
$$

Como la cantidad

$$\sum_{1\le i_1<\ldots<i_b\le n} I\left\{\min_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} < m_{i,j,k}\right\} I\left\{\max_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} > M_{i,j,k}\right\}$$

representa el número de bandas formadas por b puntos que están formadas por al menos un punto inferior al mínimo y otro superior al máximo, las combinaciones posibles, si las coordenadas son distintas, son

$$l_{m_{i,j,k}}\left(n - l_{M_{i,j,k}} - \eta_{M_{i,j,k}}\right)\binom{n-2}{b-s},$$

es decir, $l_{m_{i,j,k}}$ posibles puntos por debajo del mínimo que se combinan con $n - l_{M_{i,j,k}} - \eta_{M_{i,j,k}}$ puntos por encima del máximo y para el resto de componentes de la banda de tamaño b, que son $b-2$, se puede coger cualquiera de los demás, $n-2$. Si las coordenadas son iguales las posibilidades son

$$l_{i,k}\left(n - l_{i,k} - \eta_{i,k}\right)\binom{n-2}{b-s}.$$

El número de combinaciones posibles para

$$\sum_{1\le i_1<\ldots<i_b\le n} I\left\{\min_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} < m_{i,j,k}\right\} I\left\{\max_{t\in\{i_1,i_2,\ldots,i_b\}} x_{t,k} = M_{i,j,k}\right\}$$

será

$$\eta_{M_{i,j,k}} l_{m_{i,j,k}} \binom{l_{M_{i,j,k}} + \eta_{M_{i,j,k}} - 2}{B-2}$$

si las coordenadas son diferentes y, si son iguales,

$$\eta_{i,k} l_{i,k} \binom{l_{i,k} + \eta_{i,k} - 2}{B-2}.$$

De forma análoga, para

$$\sum_{1 \le i_1 < \ldots < i_b \le n} I\left\{ \min_{t \in \{i_1, i_2, \ldots, i_b\}} x_{t,k} = m_{i,j,k} \right\} I\left\{ \max_{t \in \{i_1, i_2, \ldots, i_b\}} x_{t,k} > M_{i,j,k} \right\}$$

será

$$\eta_{m_{i,j,k}} \left(n - l_{M_{i,j,k}} - \eta_{M_{i,j,k}} \right) \binom{n - l_{m_{i,j,k}} - 2}{b-2}$$

si las coordenadas son diferentes y, si son iguales,

$$\eta_{i,j} \left(n - l_{i,k} - \eta_{i,k} \right) \binom{n - l_{i,k} - 2}{b-2}.$$

Y finalmente, para

$$\sum_{1 \le i_1 < \ldots < i_b \le n} I\left\{ \min_{t \in \{i_1, i_2, \ldots, i_b\}} x_{t,k} = m_{i,j,k} \right\} I\left\{ \max_{t \in \{i_1, i_2, \ldots, i_b\}} x_{t,k} = M_{i,j,k} \right\}$$

se tiene, si las coordenadas son distintas,

$$\eta_{m_{i,j,k}} \eta_{M_{i,j,k}} \binom{l_{M_{i,j,k}} - l_{m_{i,j,k}} + \eta_{M_{i,j,k}} - 2}{B-2}$$

y, si son iguales,

$$\binom{\eta_{i,k}}{2} \binom{l_{i,k} - l_{i,k} + \eta_{i,k} - 2}{B-2}.$$

Por último. se multiplica el número de bandas que contienen a las dos coordenadas por $\binom{n}{b}^{-1}$ para obtener el promedio de bandas formadas por b puntos que contienen a la coordenada k-ésima de los puntos, teniéndose que

$$\begin{aligned} SBM_n^{b,k)}\left(x_{i,k}, x_{j,k}\right) \;=\; & \binom{n}{b}^{-1} \left[l_{m_{i,j,k}} \left(n - l_{M_{i,j,k}} - \eta_{M_{i,j,k}} \right) \binom{n-2}{b-s} \right. \\ & + \eta_{m_{i,j,k}} \left(n - l_{M_{i,j,k}} - \eta_{M_{i,j,k}} \right) \binom{n - l_{m_{i,j,k}} - 2}{b-2} \\ & + \eta_{M_{i,j,k}} l_{m_{i,j,k}} \binom{l_{M_{i,j,k}} + \eta_{M_{i,j,k}} - 2}{B-2} \\ & \left. + \eta_{m_{i,j,k}} \eta_{M_{i,j,k}} \binom{l_{M_{i,j,k}} - l_{m_{i,j,k}} + \eta_{M_{i,j,k}} - 2}{B-2} \right] \end{aligned}$$

si las coordenadas son distintas, y

$$\begin{aligned} SBM_n^{b,k)}\left(x_{i,k},x_{j,k}\right) &= \binom{n}{b}^{-1}\left[l_{i,k}\left(n-l_{i,k}-\eta_{i,k}\right)\binom{n-2}{b-s}\right. \\ &+\eta_{i,j}\left(n-l_{i,k}-\eta_{i,k}\right)\binom{n-l_{i,k}-2}{b-2} \\ &+\eta_{i,k}l_{i,k}\binom{l_{i,k}+\eta_{i,k}-2}{B-2} \\ &\left.+\binom{\eta_{i,k}}{2}\binom{l_{i,k}-l_{i,k}+\eta_{i,k}-2}{B-2}\right] \end{aligned}$$

si son iguales. ■

Observación 2.4 *Si en cada columna de la matriz X no se repite ningún valor, las fórmulas para el cálculo son*

$$\begin{aligned} SBM_n^{b,k)}\left(x_{i,k},x_{j,k}\right) &= l_{m_{i,j,k}}\left(n-l_{M_{i,j,k}}-1\right)\binom{n-2}{b-s} \\ &+\left(n-l_{M_{i,j,k}}-1\right)\binom{n-l_{m_{i,j,k}}-2}{b-2} \\ &+l_{m_{i,j,k}}\binom{l_{M_{i,j,k}}-1}{B-2} \\ &+\binom{l_{M_{i,j,k}}-l_{m_{i,j,k}}-1}{B-2}, \end{aligned}$$

si $i \neq j$ y

$$\begin{aligned} SBM_n^{b,k)}\left(x_{i,k},x_{j,k}\right) &= l_{i,k}\left(n-l_{i,k}-1\right)\binom{n-2}{b-s} \\ &+\left(n-l_{i,k}-1\right)\binom{n-l_{i,k}-2}{b-2} \\ &+l_{i,k}\binom{l_{i,k}-1}{B-2} \end{aligned}$$

si $i = j$.

Observación 2.5 *Si además de no haber valores repetidos en cada columna de X se realiza el cálculo para $B = 2$ entonces las fórmulas se reducen a*

$$SBM_n^{b,k)}\left(x_{i,k},x_{j,k}\right) = \left(l_{m_{i,j,k}}+1\right)\left(n-l_{M_{i,j,k}}\right)$$

si $i \neq j$ y

$$SBM_n^{b,k)}\left(x_{i,k},x_{j,k}\right) = \left(l_{i,k}+1\right)\left(n-l_{i,k}\right)-1,$$

si $i = j$.

Observación 2.6 *Si se quiere diseñar un algoritmo para el cálculo de la similaridad por bandas se recomienda comenzar creando una matriz de tamaño $n \times n$ que contenga, para las posibles combinaciones de $l_{m_{i,j,k}}$ y $l_{M_{i,j,k}}$, el número de bandas que contienen a los puntos. Posteriormente ordenar por columnas la matriz X y, para cada comparación de pares de puntos en cada dimensión, tomar el valor correspondiente de la matriz calculada previamente. De esta manera el algoritmo sólo hace cálculos una vez y no hay que explorar todas las posibles bandas, reduciéndose el tiempo de cálculo sustancialmente, sobre todo cuando la dimensión es elevada.*

2.3. Ejemplos de aplicación de las similaridades

Para mostrar la utilidad y el funcionamiento de la idea de similaridad basada en profundidad, se han simulado dos conjuntos de datos de tamaño muestral 40 y de dimensión dos, sobre los que se aplican las distintas similaridades propuestas. El primer conjunto de datos se generó a partir de la distribución normal estándar bivariante y el segundo a partir de la composición de dos distribuciones exponenciales independientes de parámetro igual a uno.

Para todas las similaridades propuestas se presentan varios gráficos. El primero es una superficie tridimensional que muestra las similaridades de los puntos del plano con respecto a otro punto que se toma fijo, calculadas para la muestra de datos normales. En segundo lugar, se presentan las curvas de nivel de las similaridades con respecto a un punto fijo, junto con los puntos de la muestra. Estas curvas muestran el comportamiento de la similaridad en dos situaciones: cuando el punto fijo es un punto central y cuando el punto fijo es un punto externo.

Guardando el orden de introducción se comienza con la similaridad de Mahalanobis.

2.3.1. Similaridad de Mahalanobis

Las curvas de nivel presentan siempre patrones elípticos debido al uso de la distancia de Mahalanobis. La forma de esas elipses dependerá de las relaciones entre las variables.

Así pues siempre se tendrá algún tipo de simetría en estas curvas, incluso cuando los datos sobre los que se estima la matriz de covarianzas no lo sean. Este comportamiento es el mismo que posee su función de profundidad.

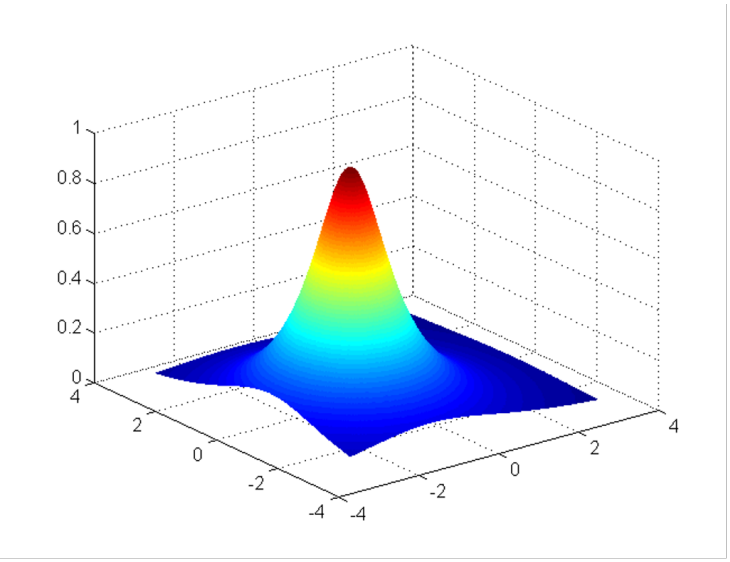
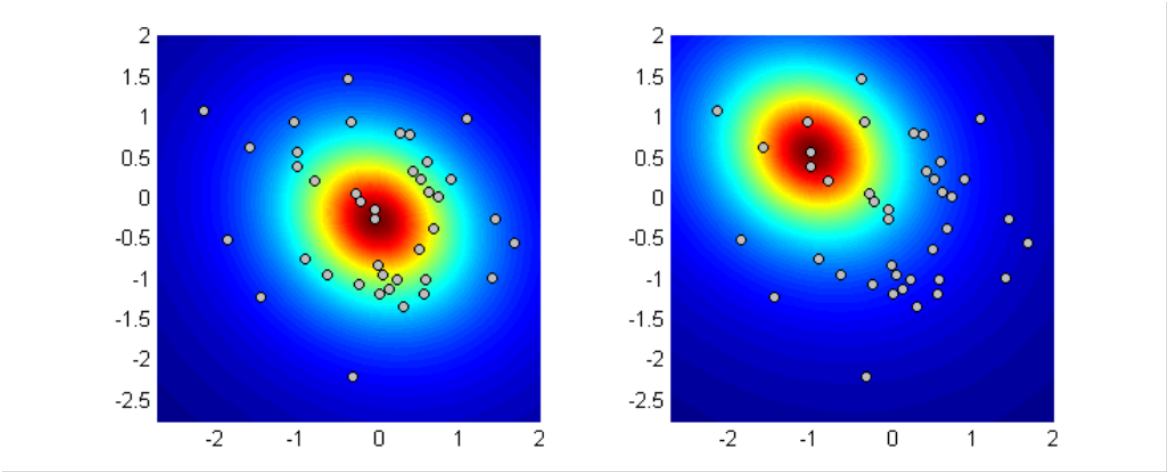


Figura 2.7: *Similaridad de Mahalanobis.*

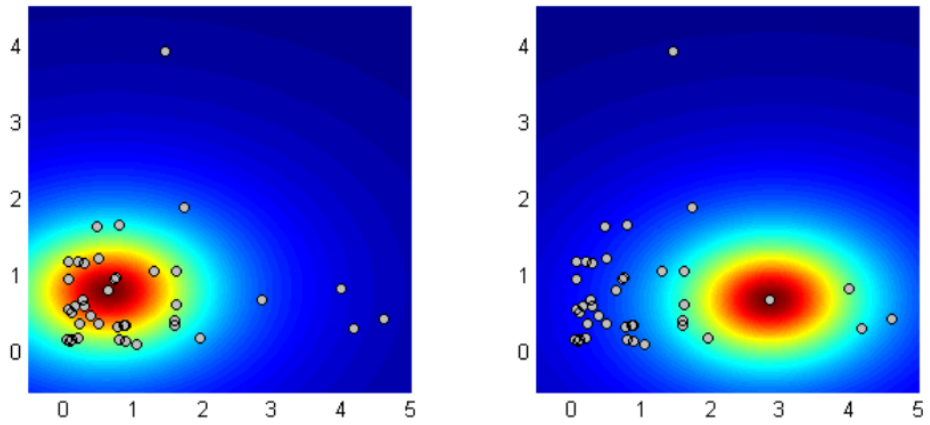
La Figura 2.7 muestra la superficie tridimensional de la similaridad fijado un punto y y para cualquier punto del espacio, x. Se puede observar la suavidad de la superficie. La Figura 2.8 muestra las curvas de nivel correspondientes a una muestra aleatoria simple de una normal estándar con dos puntos de referencia, uno central y otro externo en relación a la nube de puntos. Como puede observarse el comportamiento de las curvas de nivel es el mismo en ambos casos, no diferenciando el hecho de que un punto es más atípico o "improbable" que otro.

En cuanto a su aplicación en el caso de una distribución asimétrica (Figura 2.9), se observa de nuevo que no hay distinción alguna entre punto interno y externo; y que el hecho de que la forma de las curvas sea elíptica, en esta ocasión no refleja con fidelidad las semejanzas entre los puntos.



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.8: *Similaridad de Mahalanobis con respecto a un punto fijo para una muestra normal.*



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.9: *Similaridad de Mahalanobis con respecto a un punto fijo para una muestra exponencial.*

2.3.2. Similaridad por proyecciones

La similaridad por proyecciones, como puede observarse en las Figuras 2.10 a 2.12 es similar a la de Mahalanobis en el sentido de que la forma depende de todos los puntos pero no del punto de referencia. Este comportamiento no es el más deseable, ya que el

objetivo son medidas que tengan en cuenta la forma de la distribución y la posición de los puntos en el espacio. La capacidad de adaptación de esta similaridad para la situación asimétrica de la Figura 2.12 no se muestra demasiado elevada, aunque en este sentido mejora a la similaridad de Mahalanobis. Aun así, una ventaja que ofrece esta similaridad frente a la de Mahalanobis es la robustez, ya que las proyecciones sitúan las estimaciones en el peor de los casos, lo que produce resultados más estables.

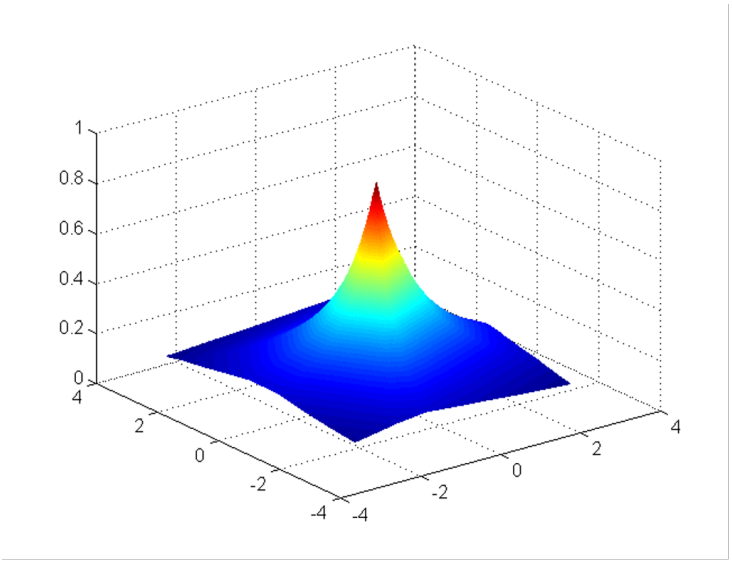
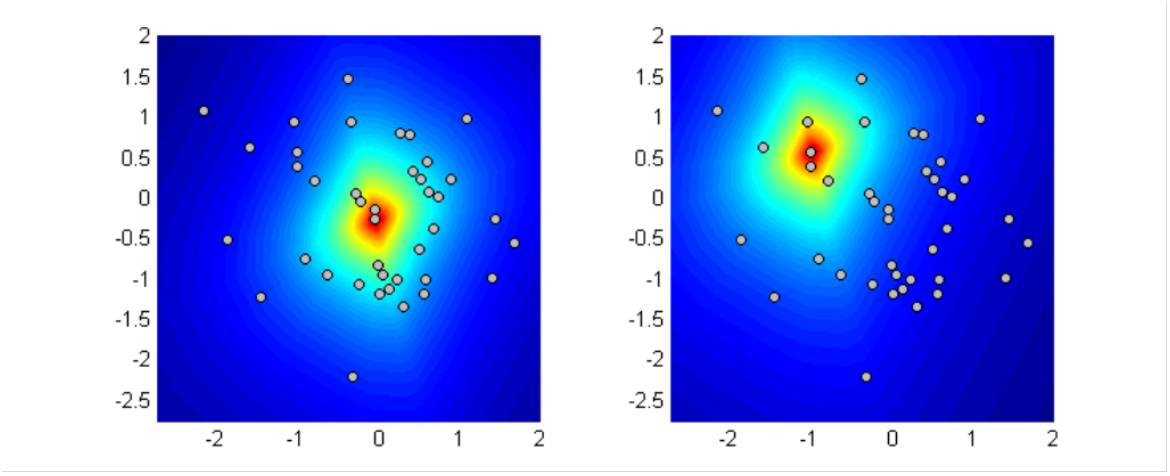


Figura 2.10: *Similaridad por proyecciones.*

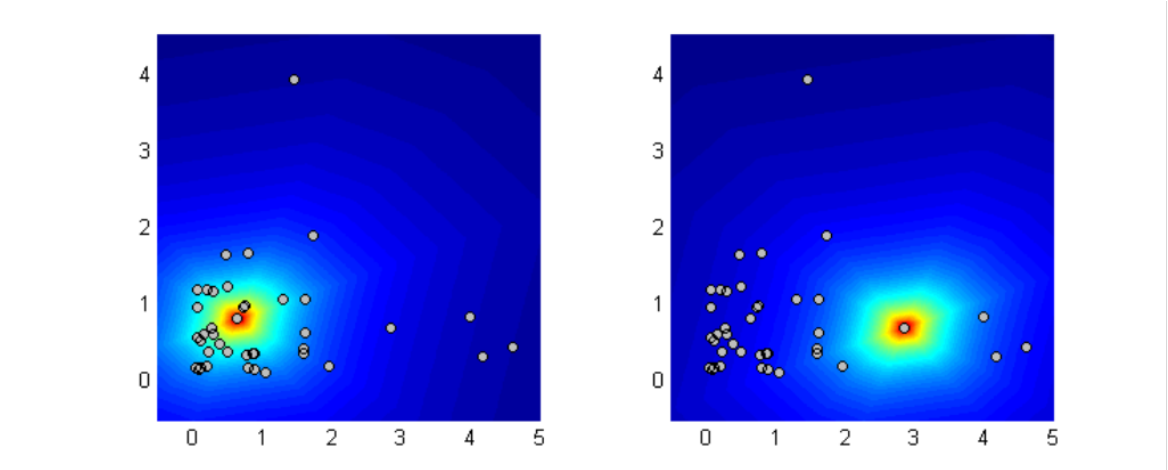
2.3.3. Similaridad de Oja

Esta similaridad presenta un comportamiento similar a la de Mahalanobis, aunque existen algunas diferencias que la hacen más atractiva que ésta. Por un lado sus curvas de nivel se adaptan mejor a los puntos de referencia y por otro lado tiene más en cuenta la forma de la distribución, produciendo similaridades mucho más adecuadas que las dos anteriores. El decrecimiento en el valor de la similaridad cuando se toma un punto alejado del de referencia es mucho más lento que en los casos anteriores. Esto puede verse en la Figura 2.13 que muestra la superficie fijado un punto. En las Figuras 2.14 y 2.15 se observa que las curvas son sustancialmente diferentes según el punto de referencia sea central o externo y según la forma de los datos sea simétrica o no.



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.11: *Similaridad por proyecciones con respecto a un punto fijo para una muestra normal.*



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.12: *Similaridad por proyecciones con respecto a un punto fijo para una muestra exponencial.*

2.3.4. Similaridad simplicial

Una de las principales características de esta similaridad es que, al igual que sucede con la profundidad simplicial, fuera de la envolvente convexa de los puntos de la muesta, la similaridad se anula. Es decir, la similaridad entre dos puntos situados fuera de dicha

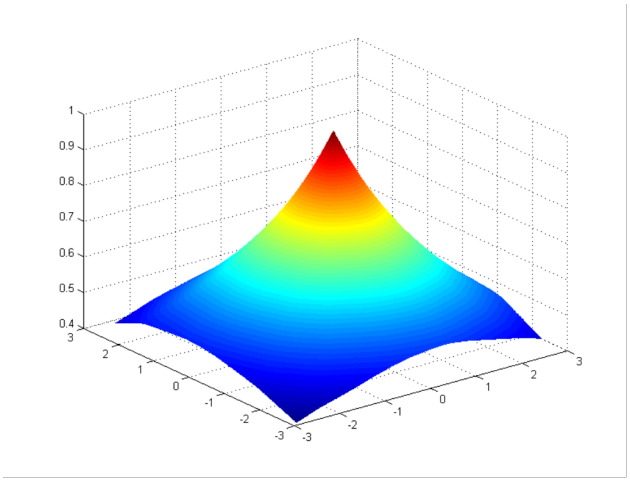
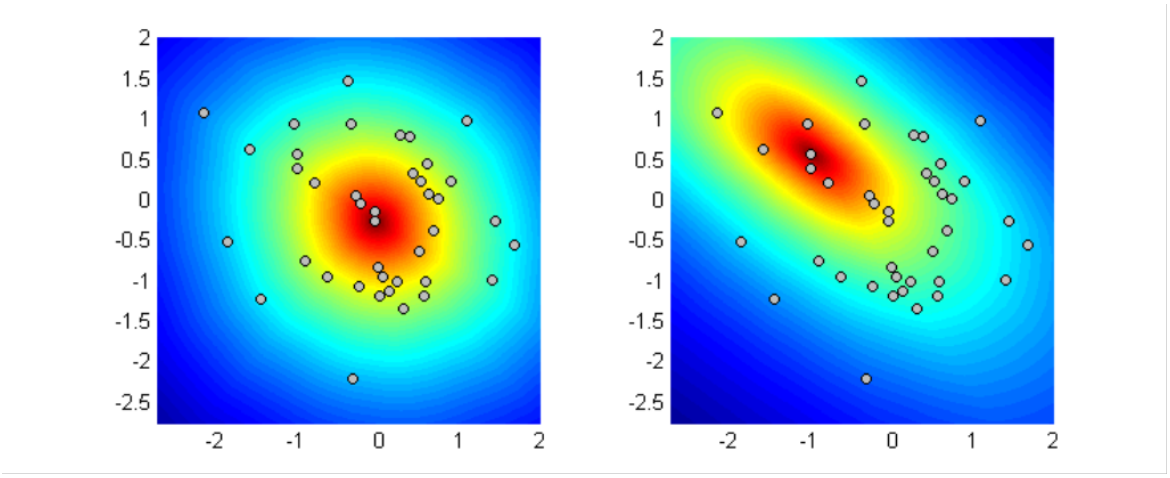


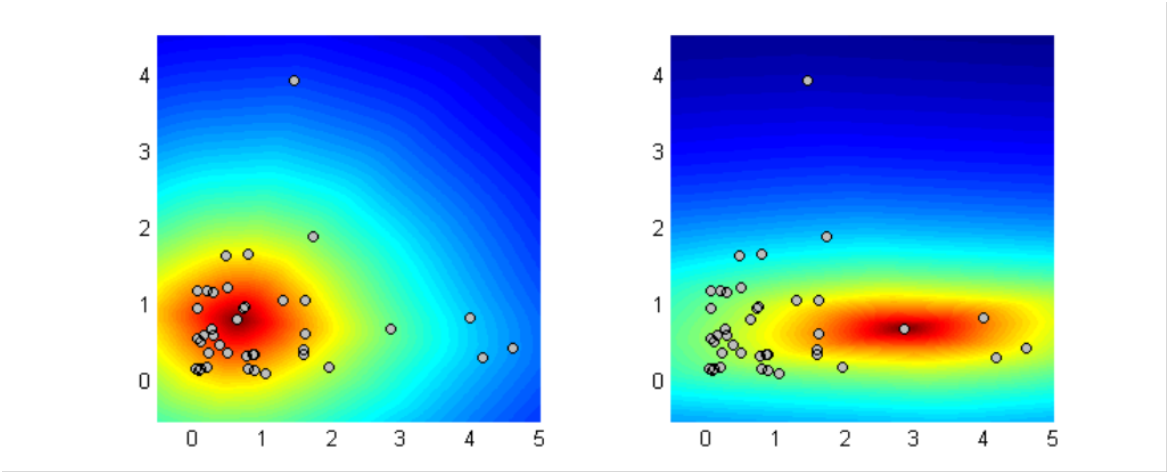
Figura 2.13: *Similaridad de Oja.*



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.14: *Similaridad de Oja con respecto a un punto fijo para una muestra normal.*

envolvente vale cero y la similaridad entre un punto de dentro y otro de fuera también. Esto puede verse como una ventaja o como un inconveniente. Como un inconveniente debido a que siempre que se trabaje con la similaridad muestral, aunque la distribución generadora sea no nula en todo el espacio, se tendrá una envolvente convexa acotada y habrá infinitos puntos (todos los de fuera de la envolvente) para los que no se estimará correctamente la similaridad. Y como ventaja, el caso en que la distribución generadora de los datos tome valores no nulos sobre conjuntos acotados de alguna manera, ya que no



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.15: *Similaridad de Oja con respecto a un punto fijo para una muestra exponencial.*

se obtendrán similaridades positivas para puntos con función de densidad nula.

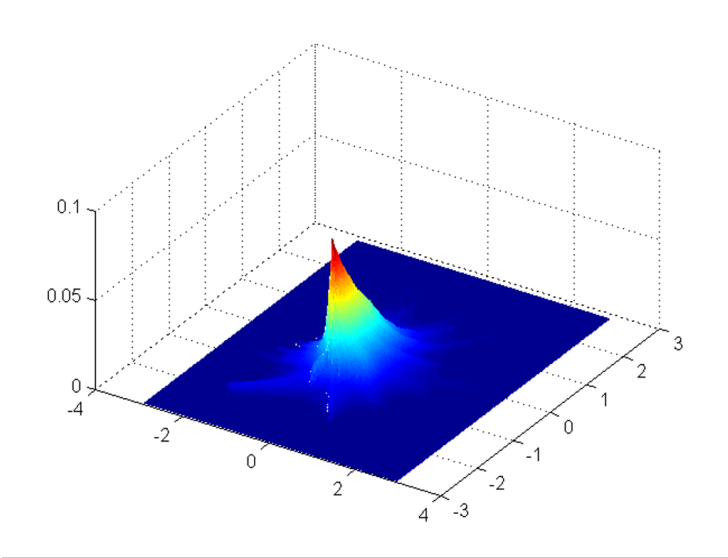
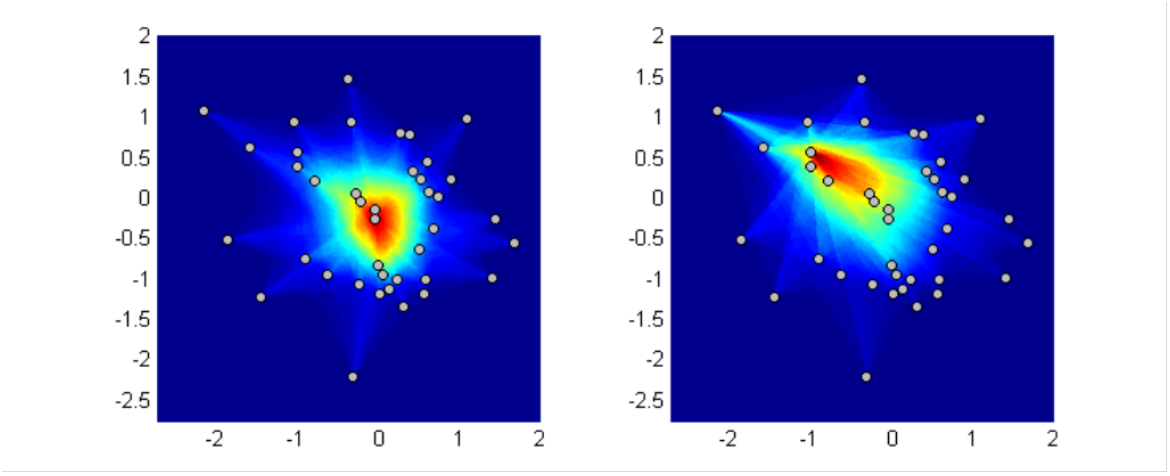


Figura 2.16: *Similaridad simplicial.*

En cuanto al comportamiento, se tiene que las curvas de nivel tienen picos en cada uno de los puntos de la muestra, lo que produce que estas curvas no sean convexas. Esto es debido a que la similaridad está compuesta por símplices (en dimensión dos, triángulos). Por otro lado, como se puede observar en la Figura 2.16, la velocidad con

que el valor de la similaridad decrece cuando uno de los puntos se aleja del otro es más rápida que en las similaridades anteriores. Además, a diferencia de éstas, presenta una capacidad de adaptación tanto a la forma de la nube de puntos como a la posición de los puntos a comparar, mucho más elevada. Esto puede observarse en las Figuras 2.17 y 2.18. Puede notarse también cómo se comporta sobre variables acotadas, ya que, en el caso exponencial (Figura 2.18), asigna valores nulos a puntos fuera del primer cuadrante.

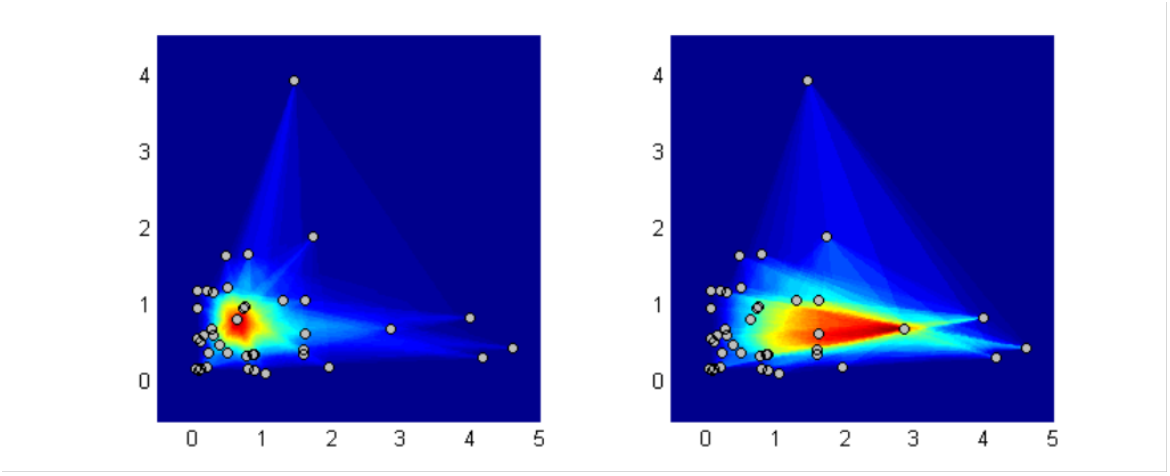


(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.17: *Similaridad simplicial con respecto a un punto fijo para una muestra normal.*

2.3.5. Similaridad por bandas

En las Figuras 2.19 a 2.21 se presentan los gráficos de la similaridad por bandas para $B = 2$ (bandas formadas por dos puntos). Esta similaridad, como puede verse en las gráficas, comparte algunas propiedades de las comentadas para la similaridad simplicial. Fuera del menor hipercubo que contiene a todos los puntos la similaridad se anula, es decir, si se comparan dos puntos de fuera o uno de fuera y uno de dentro la similaridad vale cero. Más aún, como puede observarse en las Figuras 2.20 y 2.21 existen zonas dentro de ese hipercubo mínimo cuyos puntos tienen valores iguales a cero. A pesar de esto se tiene una capacidad de adaptación razonablemente buena, tanto cuando se trata de distribuciones de formas diferentes como cuando se toman puntos de referencia centrales



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.18: *Similaridad simplicial con respecto a un punto fijo para una muestra exponencial.*

o externos.

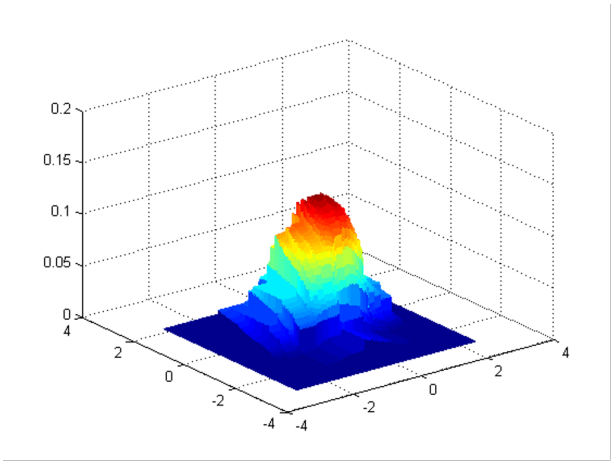
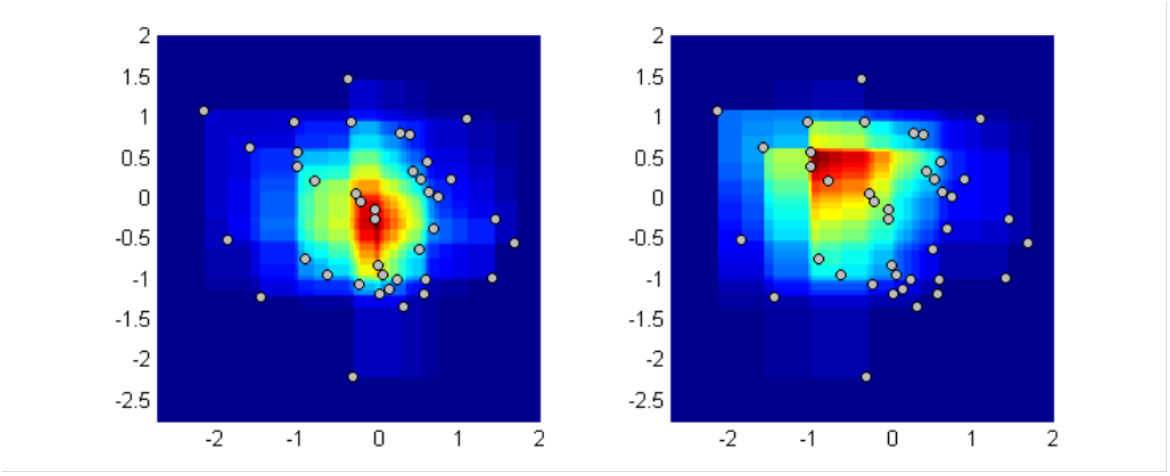


Figura 2.19: *Similaridad por bandas.*

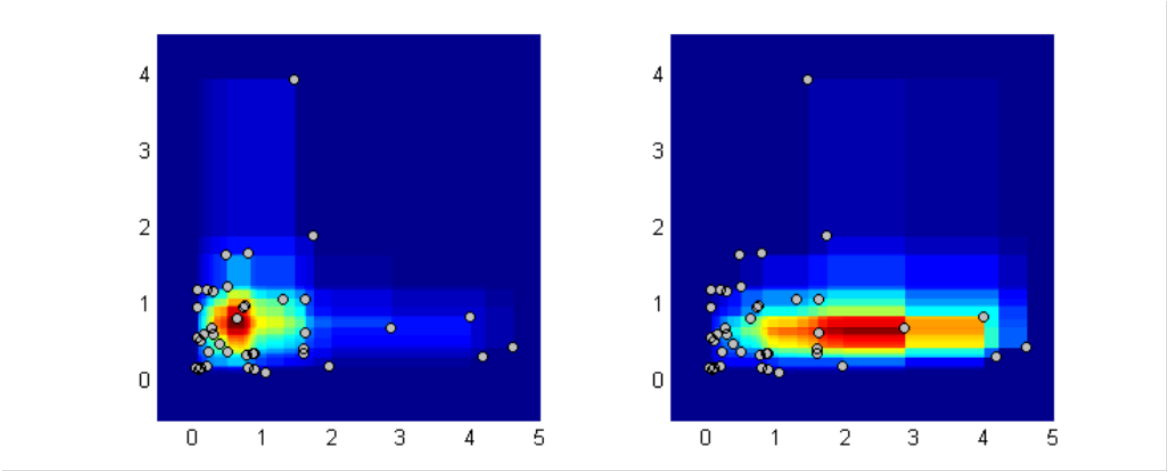
2.3.6. Similaridad por bandas modificada

Como en el caso anterior, las Figuras han sido generadas mediante bandas formadas por dos puntos. De nuevo, algunos comentarios sobre las propiedades coinciden con los de la similaridad por bandas, aunque en esta ocasión se observa claramente cómo no se



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.20: *Similaridad por bandas con respecto a un punto fijo para una muestra normal.*



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.21: *Similaridad por bandas con respecto a un punto fijo para una muestra exponencial.*

produce el desvanecimiento en el infinito. La Figura 2.22 muestra claramente la superficie en la que aparece la cruz que marca las direcciones sobre las que no se anula la similaridad (direcciones paralelas a los ejes x e y). La similaridad se anula para cuadrantes con origen los cuatro vértices del mínimo hipercubo que contiene a todas las observaciones: puntos con la primera coordenada mayor que el máximo en esa coordenada y con la segunda

mayor que el máximo de las segundas; primera coordenada mayor que el máximo de la primera y segunda menor que el mínimo de la segunda; primera coordenada menor que el mínimo de la primera y la segunda menor que el mínimo de la segunda; y, por último, los que tienen la primera coordenada menor que el mínimo de la primera y segunda mayor que el máximo de la segunda. En cuanto a la capacidad de adaptación de esta similaridad, Figuras 2.23 y 2.24, se tiene un resultado muy satisfactorio dentro del hipercubo mínimo. Fuera de él, los resultados no parecen fiables.

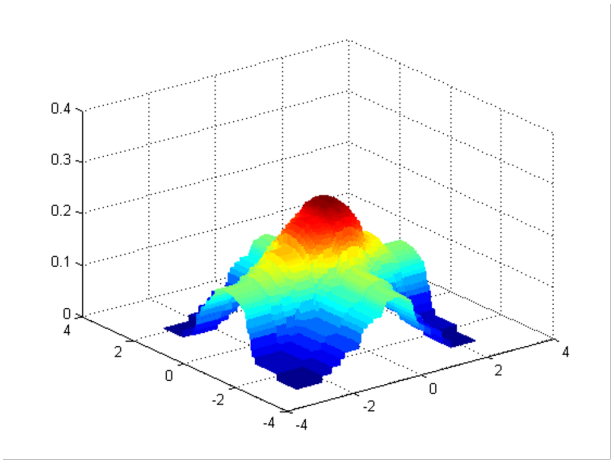
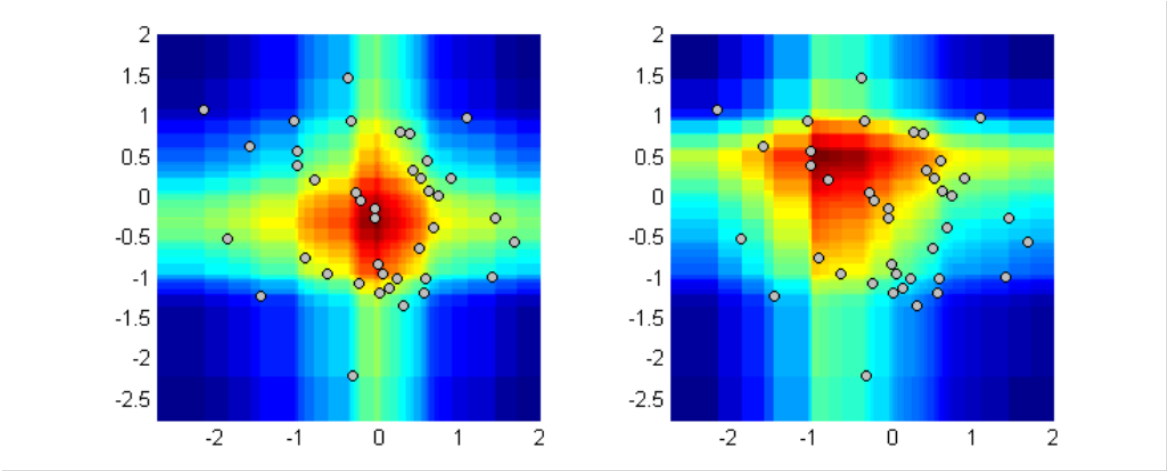


Figura 2.22: *Similaridad por bandas modificada.*

2.4. Propiedades de las similaridades

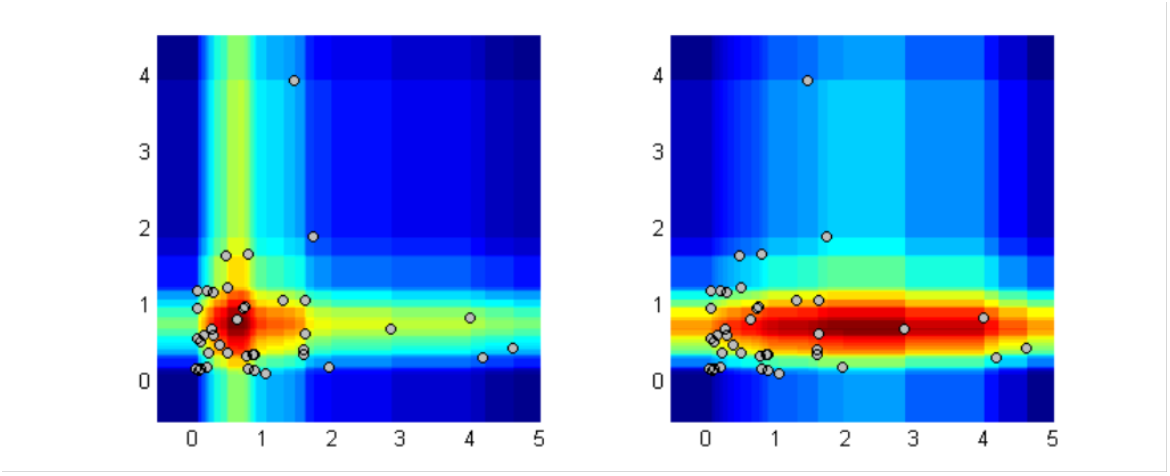
En esta sección se estudian algunas propiedades de las similaridades introducidas anteriormente. En Liu (1990) y Zuo y Serfling (2000a) se proponen varias propiedades que las funciones de profundidad deben cumplir para asegurar que las ordenaciones y puntuaciones que asignen sean congruentes. Estas propiedades se adaptan aquí a las similaridades basadas en profundidad con el fin de asegurar que las proximidades calculadas a partir de éstas sean adecuadas.

Por otro lado, también se analiza tanto el comportamiento asintótico de las versiones muestrales de las similaridades como la propiedad de continuidad. Las propiedades que debe verificar una función para ser una similaridad se incluyen también en esta sección. Se



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.23: *Similaridad por bandas modificada con respecto a un punto fijo para una muestra normal.*



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 2.24: *Similaridad por bandas modificada con respecto a un punto fijo para una muestra exponencial.*

comienza con las propiedades deseables como funciones basadas en la idea de profundidad.

2.4.1. Propiedades como funciones basadas en profundidad

Las propiedades deseables para las funciones de profundidad que fueron propuestas en Liu (1990) y Zuo y Serfling (2000a) son que la función tiene que alcanzar su máximo valor en el centro de la distribución (si ésta lo tiene), que sobre cualquier recta con origen el centro de la distribución la función debe ser monótona decreciente, que su límite cuando los puntos se alejan del centro tiene que ser cero y que ha de ser invariante ante transformaciones afines.

Las propiedades que se proponen y estudian para las similaridades basadas en profundidad son adaptaciones de estas propiedades más la inclusión de la propiedad de simetría que todas las similaridades propuestas cumplen por definición. A continuación se enumeran estas propiedades de las similaridades:

1. Al comparar un punto x con cualquier punto y del espacio, el valor máximo de la similaridad entre x e y es igual al de la similaridad entre x y él mismo.

2. Al comparar un punto x con cualquier punto y del espacio, la similaridad entre el punto x y entre cualquier otro punto del segmento que une x e y es mayor o igual que la que hay entre x e y.

3. La similaridad entre los puntos x e y tiende a cero al alejarse el punto y de x.

4. La similaridad entre x e y es igual a la similaridad entre transformaciones afines de estos puntos (y de su distribución de referencia).

5. La similaridad tiene que ser simétrica.

Si la similaridad verifica estas cinco propiedades se dice similaridad por profundidad.

Definición 2.7 *Sea F_X una función de distribución d-dimensional. La función acotada y no negativa $S(x, y; F_X)$ se llama similaridad basada en profundidad si verifica*

(i) $S(y, y; F_X) = \sup_{x \in \mathbb{R}^d} S(x, y; F_X)$, *para cualquier* $y \in \mathbb{R}^d$

(ii) Para cualesquiera $x, y \in \mathbb{R}^d$ *y para todo* $\alpha \in [0,1]$ *se tiene que* $S(x, y; F_X) \leq S(y + \alpha(x - y), y; F_X)$

(iii) Para cualquier $y \in \mathbb{R}^d$ *se tiene que* $S(x, y; F_X) \to 0$ *cuando* $\|x\| \to \infty$

(iv) $S(x, y; F_X) = S(Ax + b, Ay + b; F_{AX+b})$ *para cualquier par de vectores* x *e* y *en* $\mathbb{R}^d$*, cualquier matriz* A *no singular de tamaño* $d \times d$ *y cualquier vector* $b \in \mathbb{R}^d$

(v) Para cualquier par de puntos x *e* y *en* $\mathbb{R}^d$*, se cumple que* $S(x, y; F_X) = S(y, x; F_X)$

Fijado un punto x en $\mathbb{R}^d$, si se aplica la función a un conjunto de puntos y se ordenan los valores de mayor a menor se obtiene una ordenación de los puntos de más próximos a más alejados de x.

Proposición 2.1 *Si* Σ_F *es equivariante afín, es decir,* $\Sigma_{F(AX)} = A\Sigma_F A'$*, entonces la similaridad de Mahalanobis es una similaridad basada en profundidad en el sentido de la Definición 2.7.*

Demostración. La demostración de las propiedades es inmediata:

(i) Para cualquier punto $y \in \mathbb{R}^d$, $\sup_{x \in \mathbb{R}^d} SM(x, y; F) = 1$ y ese valor se alcanza cuando la distancia de Mahalanobis es cero:

$$d^2_{Mah}(x, y) = (x - y)' \Sigma^{-1} (x - y) = 0 \Leftrightarrow x = y$$

por lo tanto, el máximo de la función $SM(x, y; F)$ fijado el punto y se obtiene para $x = y$.

(ii) El decrecimiento monótono también se verifica. Sean x e $y \in \mathbb{R}^d$, y $\alpha \in [0,1]$; entonces

$$SM(\alpha x + (1 - \alpha) y, y; F) \geq SM(x, y; F) \Longleftrightarrow$$

$$\Longleftrightarrow d^2_{Mah}(\alpha x + (1 - \alpha) y, y) \leq d^2_{Mah}(x, y)$$

$$\begin{aligned} d^2_{Mah}\left(\alpha x+(1-\alpha)\, y, y\right) &= \left(\alpha x+(1-\alpha)\, y-y\right)' \Sigma^{-1}\left(\alpha x+(1-\alpha)\, y-y\right) \\ &= \left(\alpha x-\alpha y\right)' \Sigma^{-1}\left(\alpha x-\alpha y\right) \\ &= \alpha^2 \left(x-y\right)' \Sigma^{-1}\left(x-y\right) \\ &\underset{\alpha^2\in[0,1]}{\leq} \left(x-y\right)' \Sigma^{-1}\left(x-y\right) = d^2_{Mah}\left(x,y\right) \end{aligned}$$

(iii) Se cumple, ya que

$$\lim_{\|x\|\to\infty} d^2_{Mah}\left(x,y\right)=\infty$$

(iv) La invarianza afín se verifica ya que la distancia de Mahalanobis es afín invariante:

$$\begin{aligned} d^2_{Mah,A\Sigma A'}\left(Ax+b, Ay+b\right) &= \left(Ax+b-\left(Ay+b\right)\right)' \left(A\Sigma A'\right)^{-1}\left(Ax+b-\left(Ay+b\right)\right) \\ &= \left(Ax-Ay\right)' \left(A'\right)^{-1} \Sigma A^{-1}\left(Ax-Ay\right) \\ &= \left(x-y\right)' A' \left(A'\right)^{-1} \Sigma A^{-1} A \left(x-y\right) \\ &= \left(x-y\right)' \Sigma \left(x-y\right) = d^2_{Mah}\left(x,y\right), \end{aligned}$$

(v) La simetría es cierta ya que la distancia de Mahalanobis también lo es. ■

Proposición 2.2 *La similaridad por proyecciones es una similaridad basada en profundidad en el sentido de la Definición 2.7.*

Demostración. Se comprueba cada una de las propiedades:

(i) La función $SP\left(x,y;F\right)$, fijado un punto y, alcanza su máximo para valores de x, cuando la función de atipicidad $A\left(x,y;F\right)$ toma su valor mínimo. Lo que ocurre cuando $x=y$

$$0\leq \min_{x\in\mathbb{R}^d} A\left(x,y;F\right)=\min_{x\in\mathbb{R}^d}\sup_{\|u\|=1}\frac{|u'x-u'y|}{Meda\left(u'X\right)}\leq \sup_{\|u\|=1}\left.\frac{|u'x-u'y|}{Meda\left(u'X\right)}\right|_{x=y}=0.$$

(ii) El decrecimiento monótono se verifica ya que

$$|u'\left(\alpha x+(1-\alpha)\, y\right)-u'y|=|\alpha u'x-\alpha u'y|=|\alpha|\,|u'x-u'y|\leq |u'x-u'y|$$

y, por lo tanto,

$$\frac{|u'\left(\alpha x+(1-\alpha)\, y\right)-u'y|}{Meda\left(u'X\right)}\leq \frac{|u'x-u'y|}{Meda\left(u'X\right)},$$

lo que implica que

$$\sup_{\|u\|=1} \frac{|u'(\alpha x + (1-\alpha)y) - u'y|}{Meda(u'X)} \leq \sup_{\|u\|=1} \frac{|u'x - u'y|}{Meda(u'X)}$$

y

$$SP(\alpha x + (1-\alpha)y, y; F) \geq SP(x, y; F).$$

(iii) Se verifica debido a que, cuanto más alejados estén los puntos x e y, más alejadas estarán sus proyecciones y por lo tanto mayor será $A(x, y; F)$. En el límite la función será infinita.

(iv) La similaridad es afín invariante ya que $A(x, y; F)$ lo es también.

(v) Igual ocurre con la simetría, $A(x, y; F)$ es simétrica. ■

Proposición 2.3 *La similaridad de Oja verifica todas las propiedades de la Definición 2.7, excepto la de invarianza afín.*

Demostración. Se demuestra cada una de las cuatro propiedades que verifica:

(i) Fijado un punto y, el máximo de la función $SO(x, y; F)$ se obtiene cuando $x = y$, ya que

$$\max_{x\in\mathbb{R}^d} SO(x, y; F) = \max_{x\in\mathbb{R}^d} [1 + E(Vol(S[x, y, X_1, X_2, \ldots, X_{d-1}]))]^{-1}$$

es equivalente a

$$\min_{x\in\mathbb{R}^d} E(Vol(S[x, y, X_1, X_2, \ldots, X_{d-1}]))$$

y el volumen de todos los símplices aleatorios es igual a cero sólo en el caso en que dos de los vertices coincidan, es decir, se tiene que dar que $x = y$. En ese caso $SO(x, y; F) = 1$.

(iii) En cuanto al decrecimiento monótono, dados x e y, como

$$SO(x, y; F) \leq SO(\alpha x + (1-\alpha)y, y; F)$$

es equivalente a

$$E\left(Vol(S[x,y,X_1,X_2,\ldots,X_{d-1}])\right) \geq E\left(Vol(S[\alpha x+(1-\alpha)\,y,y,X_1,X_2,\ldots,X_{d-1}])\right),$$

basta con probar esa desigualdad. Dada una muestra aleatoria de tamaño $d-1$ de F, denotada por $x_1, x_2, \ldots, x_{d-1}$, el volumen del símplice para x y $\alpha x + (1-\alpha)\,y$ verifica que

$$Vol(S[x,y,x_1,x_2,\ldots,x_{d-1}]) \geq Vol(S[\alpha x + (1-\alpha)\,y, y, x_1, x_2, \ldots, x_{d-1}])$$

ya que

$$\begin{aligned}
& Vol(S[\alpha x + (1-\alpha)\,y, y, x_1, x_2, \ldots, x_{d-1}]) \\
= \quad & \frac{1}{(d+1)!}\left|\det\begin{pmatrix} \alpha x + (1-\alpha)\,y & y & x_1 & \ldots & x_{d-1} \\ 1 & 1 & 1 & \ldots & 1 \end{pmatrix}\right| \\
\underset{C_1 = C_1 - (1-\alpha)C_2}{=} \quad & \frac{1}{(d+1)!}\left|\det\begin{pmatrix} \alpha x & y & x_1 & \ldots & x_{d-1} \\ \alpha & 1 & 1 & \ldots & 1 \end{pmatrix}\right| \\
= \quad & \frac{1}{(d+1)!}\,|\alpha|\left|\det\begin{pmatrix} x & y & x_1 & \ldots & x_{d-1} \\ 1 & 1 & 1 & \ldots & 1 \end{pmatrix}\right| \\
= \quad & |\alpha|\,Vol\left(S\left[x,y,x_1,x_2,\ldots,x_{d-1}\right]\right) \\
\underset{|\alpha|\leq 1}{\leq} \quad & Vol\left(S\left[x,y,x_1,x_2,\ldots,x_{d-1}\right]\right).
\end{aligned}$$

Y como esto es cierto para cualquier la muestra aleatoria de tamaño $d-1$, entonces su valor esperado también lo verifica.

(iv) El volumen es una función no acotada, por tanto si uno de los puntos está muy alejado del otro, el volumen del símplice puede tomar un valor arbitrariamente grande y la esperanza sobre todos los posibles símplices también puede ser tan grande como se desee. Por tanto, $SO\,(x,y;F)$ tiende a cero.

(v) La simetría se verifica por definición. ■

La definición de similaridad de Oja propuesta no verifica la propiedad de invarianza afín, pero puede realizarse una modificación para resolver esta carencia y obtener una similaridad que cumpla con todas las propiedades.

Proposición 2.4 *La similaridad simplicial es una similaridad basada en profundidad en el sentido de la Definición 2.7 para funciones absolutamente continuas.*

Demostración. En la demostración se hace uso, tanto de las propiedades que la similaridad hereda de la profundidad simplicial, como de los conjuntos que se definen a continuación. Dados dos puntos x e y en $\mathbb{R}^d$, se definen los sucesos A, B, C y A_α como conjuntos de símplices aleatorios que verifican determinadas condiciones de pertenencia de puntos x e y y combinaciones lineales convexas suyas. Se definen

$$\begin{aligned} A &= \{X_1, X_2, \ldots, X_{d+1} : x, y \in S[X_1, X_2, \ldots, X_{d+1}]\}, \\ B &= \{X_1, X_2, \ldots, X_{d+1} : x \in S[X_1, X_2, \ldots, X_{d+1}]\}, \\ C &= \{X_1, X_2, \ldots, X_{d+1} : y \in S[X_1, X_2, \ldots, X_{d+1}]\} \text{ y} \\ A_\alpha &= \{X_1, X_2, \ldots, X_{d+1} : \alpha x + (1-\alpha) y, y \in S[X_1, X_2, \ldots, X_{d+1}]\}, \alpha \geq 0. \end{aligned}$$

Estos sucesos verifican que $A \subseteq B$ y $A \subseteq C$ y, debido a la convexidad de los símplices, que si $\alpha_1 \geq \alpha_2$ entonces $A_{\alpha_1} \subseteq A_{\alpha_2}$. Se demuestra cada una de las propiedades:

(i) Fijado un punto y, el máximo de $SS(x, y; F)$ se alcanza cuando $x = y$, ya que $SS(x, y; F) = Pr(A) \leq Pr(C) = SS(y, y; F)$

(ii) La propiedad de invarianza afín se justifica gracias a la convexidad de los símplices, es decir, dada la matriz no singular A y el vector b, se tiene que $x \in S[x_1, x_2, \ldots, x_{d+1}]$ es equivalente a que $Ax + b \in S[Ax_1 + b, Ax_2 + b, \ldots, Ax_{d+1} + b]$ y, por tanto, la probabilidad de que eso ocurra para todos los símplices aleatorios es la misma.

(iii) El decrecimiento monótono se asegura también por la convexidad del conjunto $S[X_1, X_2, \ldots, X_{d+1}]$, ya que todos los símplices que contienen a los puntos x e y, por convexidad, contienen también a los todos los puntos que son combinaciones lineales, $\alpha x + (1-\alpha) y$, por lo tanto, los conjuntos A y A_α definidos arriba cumplen que $A \subseteq A_\alpha$ y, por consiguiente, $SS(x, y; F) = Pr(A) \leq Pr(A_\alpha) = SS(\alpha x + (1-\alpha) y, y; F)$.

(iv) El desvanecimiento en el infinito se obtiene fácilmente ya que esto se verifica para la profundidad simplicial

$$SS\left(x,y;F\right)=Pr\left(A\right)\leq Pr\left(B\right)=SS\left(x,x;F\right)=PS\left(x;F\right).$$

Entonces, como la similaridad simplicial entre dos puntos está acotada por la profundidad de cada uno de los puntos, haciendo uso del desvanecimiento en el infinito de la profundidad simplicial (véase el Teorema 1 de Liu (1990)) se concluye que

$$0\leq \lim_{\|x\|\to\infty} SS\left(x,y;F\right)\leq \lim_{\|x\|\to\infty} PS\left(x;F\right)=0.$$

(v) La función es simétrica por definición. ■

Proposición 2.5 *La similaridad por bandas verifica todas las propiedades para ser similaridad basada en profundidad en el sentido de la definición 2.7, excepto la propiedad de invarianza afín.*

Demostración. En este caso hay que tener en cuenta que es el resultado de $B-1$ sumandos. Para cada sumando se cumplen todas las propiedades (salvo la de invarianza), por lo tanto su suma también las cumplirá. La demostración para cada sumando es análoga a la de la similaridad simplicial. La simetría se verifica por definición. La maximalidad se obtiene cuando los dos puntos son iguales, ya que el conjunto de todos los hipercubos que contienen a ambos puntos está contenido en el conjunto de todos los hipercubos que contienen a uno de ellos. La convexidad de los hipercubos asegura que dados x, y en $\mathbb{R}^d$, cualquier combinación lineal convexa z de estos puntos está también en el hipercubo, entonces los hipercubos que contienen a x e y, también contienen a z. El desvanecimiento en el infinito también se verifica debido a que la profundidad por bandas lo cumple y a que las similaridades entre dos puntos están acotadas por las profundidades de ambos puntos. ■

Proposición 2.6 *La similaridad por bandas modificada verifica todas las propiedades de la Definición 2.7 salvo la de invarianza y desvanecimiento en el infinito.*

Demostración. Como en el caso anterior hay que estudiar cada sumando. En este caso para cada dimensión se tienen intervalos, por lo tanto, debido a que son convexos, la demostración es análoga a los casos anteriores. La simetría se cumple por definición. El desvanecimiento en el infinito no se verifica ya que si una de las coordenadas de un punto está en el rango de la coordenada de la función de distribución, siempre habrá bandas que contengan a dicha coordenada por muy alejado que esté el punto con respecto a las demás coordenadas. ■

2.4.2. Propiedades de continuidad y asintóticas

A continuación se presentan y demuestran algunas propiedades de utilidad para las similaridades de la sección anterior, como son la continuidad de la similaridad cuando la función de distribución F lo es y la convergencia de las versiones muestrales cuando el tamaño muestral aumenta.

La continuidad para las similaridades de Mahalanobis, por proyecciones y Oja está garantizada ya que están basadas en distancias y medidas de atipicidad que sí lo son.

A continuación se estudia la continuidad y el comportamiento asintótico para las otras tres similaridades. En la demostración de los resultados de convergencia se hace uso del Lema 3 que aparece en Liu (1990) y que se enuncia a continuación.

Lema 2.1 (Lema 3 en Liu (1990)) *Sean F una función de distribución en $\mathbb{R}^d$ y $x_1, x_2, \ldots, x_n$ una muestra aleatoria simple de F. Sea $U_n = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \ldots < i_m \leq n} h\left(x_{i_1}, x_{i_2}, \ldots, x_{i_m}\right)$ un U-estadístico con núcleo $h(\cdot)$ de grado m. Si h está acotada, por ejemplo por c, entonces para cualquier $r \geq 2$,*

$$E\left[\left(U_n - E\left(U_n\right)\right)^r\right] \leq \frac{K}{n^{r/2}},$$

donde K es una constante que depende de c.

Se comienza con el resultado de continuidad para la similaridad simplicial.

Teorema 2.2 *Dado $y \in \mathbb{R}^d$, si F es una función de distribución absolutamente continua, entonces $SS(\cdot, y; F)$ es continua.*

Demostración. La función $SS(\cdot, y; F)$ será continua si, dada una sucesión $x_n \in \mathbb{R}^d$ que converge a x entonces el $\lim_{n\to\infty} |SS(x, y; F) - SS(x_n, y, F)| = 0$. Para probar esto se definen los siguientes sucesos

$$A_x = \{X_1, X_2, \ldots, X_{d+1} : x \in S[X_1, X_2, \ldots, X_{d+1}]\},$$

$$A_{x_n} = \{X_1, X_2, \ldots, X_{d+1} : x_n \in S[X_1, X_2, \ldots, X_{d+1}]\} \text{ y}$$

$$A_y = \{X_1, X_2, \ldots, X_{d+1} : y \in S[X_1, X_2, \ldots, X_{d+1}]\}.$$

La similaridad simplicial $SS(x, y; F)$ es igual a $Pr(A_x \cap A_y)$; por lo tanto, la diferencia $SS(x, y; F) - SS(x_n, y, F)$ es igual a $Pr(A_x \cap A_y) - Pr(A_{x_n} \cap A_y)$. Es posible acotar esta cantidad ya que

$$A_x \cap A_y \subseteq (A_{x_n} \cap A_y) \cup \left(A_x \cap \overline{A_{x_n}} \cap A_y\right),$$

teniéndose que

$$Pr(A_x \cap A_y) \leq Pr(A_{x_n} \cap A_y) + Pr\left(A_x \cap \overline{A_{x_n}} \cap A_y\right)$$

y

$$Pr(A_x \cap A_y) - Pr(A_{x_n} \cap A_y) \leq Pr\left(A_x \cap \overline{A_{x_n}} \cap A_y\right).$$

Haciendo lo mismo para $A_{x_n} \cap A_y$, se obtiene que

$$Pr(A_{x_n} \cap A_y) - Pr(A_x \cap A_y) \leq Pr\left(\overline{A_x} \cap A_{x_n} \cap A_y\right).$$

Y, por lo tanto, se tiene que

$$\begin{aligned} |Pr(A_x \cap A_y) - Pr(A_{x_n} \cap A_y)| &\leq Pr\left(A_x \cap \overline{A_{x_n}} \cap A_y\right) + Pr\left(\overline{A_x} \cap A_{x_n} \cap A_y\right) \\ &\leq Pr\left(A_x \cap \overline{A_{x_n}}\right) + Pr\left(\overline{A_x} \cap A_{x_n}\right) \\ &\leq (d+1)\, Pr(B_n), \end{aligned}$$

donde $B_n = \{X_1, X_2, \ldots, X_d : H(X_1, X_2, \ldots, X_d)$ interseca con el segmento que une x_n con $x\}$ y $H(X_1, X_2, \ldots, X_d)$ es el hiperplano que contiene a los puntos $X_1, X_2, \ldots, X_d$, ya que el suceso $\left(A_x \cap \overline{A_{x_n}}\right) \cup \left(\overline{A_x} \cap A_{x_n}\right)$ está contenido en el suceso definido como el

conjunto de triángulos con una arista cortando al segmento que une a x con x_n. Nótese además que el $\lim\sup_{n\to\infty} B_n = \{X_1, X_2, \ldots, X_d : x \in H(X_1, X_2, \ldots, X_d)\}$ es el haz de hiperplanos que contienen a x, y que es un conjunto de medida nula si F es contínua. Por último, debido a la continuidad de F, se tiene que

$$\begin{aligned}
\lim_{n\to\infty} |SS(x,y;F) - SS(x_n,y,F)| &= \limsup_{n\to\infty} |Pr(A_x \cap A_y) - Pr(A_{x_n} \cap A_y)| \\
&\leq (d+1)\limsup_{n\to\infty} Pr(B_n) \\
&\leq (d+1) Pr\left(\limsup_{n\to\infty} B_n\right) \\
&= 0. \blacksquare
\end{aligned}$$

Teorema 2.3 *Sea $D \subseteq \mathbb{R}^d$ un conjunto abierto y $F : D \longrightarrow \mathbb{R}^+$ una función de distribución absolutamente continua. Entonces la similaridad simplicial verifica que*

$$SS(x,y;F) = SS(y,y;F) \text{ si, y sólo si, } x = y.$$

Demostración. Si $x = y$, por definición se tiene que la afirmación es cierta. Basta con verificar la implicación en sentido opuesto, es decir, que $x \neq y$ implica la desigualdad $SS(x,y;F) \neq SS(y,y;F)$ (o más exactamente $SS(x,y;F) < SS(y,y;F)$).

Sean A_{xy} y A_y dos conjuntos definidos de igual forma que en la demostración anterior, entonces se tiene que $Pr(A_{xy}) \leq Pr(A_y)\ \forall x, y \in D$; por lo tanto, para terminar la demostración, es necesario probar que la probabilidad del conjunto diferencia entre ambos, $A_y \backslash A_{xy}$, es mayor que cero. Para probar esta afirmación, por simplicidad y sin pérdida de generalidad, se toma $d = 2$. La idea de la demostración es, mediante el uso de gráficos, encontrar un conjunto perteneciente a dicha diferencia $A_y \backslash A_{xy}$ y que tenga probabilidad no nula. Al suponer que el espacio es de dimensión dos se tiene que los símplices son triángulos. Los triángulos que son de utilidad para la prueba son aquellos que contienen al punto y pero no a x.

Para la elección del primer vértice del triángulo se toma un punto cualquiera dentro del conjunto D y que no esté en la semirecta con origen x y dirección $x - y$, es decir, cualquier punto de $D \backslash SL_{x,y}$, donde $SL_{x,y} = \{x + \alpha(x - y) : \alpha \geq 0\}$. Pero, como F es continua, el conjunto $SL_{x,y}$ tiene probabilidad nula y, por tanto, el espacio sobre el que se puede elegir

el primer vértice tiene probabilidad 1. Dado $X_1 \in D \backslash SL_{x,y}$, se define el conjunto S_{X_1} como $\left\{ z \in D : \overrightarrow{X_1 z} \text{ interseca con } \overrightarrow{xy} \right\}$, donde $\overrightarrow{ab}$ es el segmento que une a con b. Como puede verse en la Figura 2.25, la región $S_{X_1} \neq \phi$, por lo que tiene probabilidad mayor que cero. Más formalmente, esta afirmación sobre la probabilidad está basada en la continuidad y positividad de F y, a que, dado que D es abierto, se tiene que $\exists \varepsilon > 0 \backslash Bola\,(y, \varepsilon) \subset D$.

Dados dos vértices del triángulo, $X_1 \in D \backslash SL_{x,y}$ y $X_2 \in S_{X_1}$, la región de puntos para el

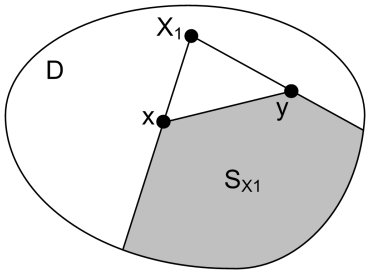


Figura 2.25: *Región para la elección del segundo punto del triángulo.*

tercer vértice (para formar un trángulo que no contenga a x, pero sí a y) puede definirse como $S_{X_1,X_2} = \{z : y \in S\,[X_1, X_2, z]\}$. Esta región está representada en la Figura 2.26 donde se representa un triángulo cualquiera que contiene a y y no a x. De nuevo, debido a que D es abierto y F es continua y positiva, se concluye que $Pr\,(S_{X_1,X_2}) > 0$ y, por último, tomando esperanzas sobre esa cantidad se tiene que $Pr\,(A_y \backslash A_{xy}) > 0$. ■

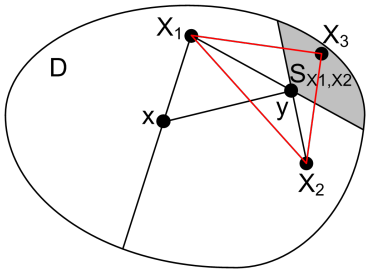


Figura 2.26: *Región S_{X_1,X_2} y triángulo perteneciente a $A_y \backslash A_{xy}$.*

Teorema 2.4 *La similaridad simplicial muestral es insesgada y consistente,*

$$SS_n\,(x, y) \xrightarrow[n\to\infty]{p} SS\,(x, y; F)\,, \forall x, y \in \mathbb{R}^d.$$

Demostración. Dados dos puntos cualesquiera x e $y \in \mathbb{R}^d$, como $SS_n(x,y)$ es un U-estadístico, se cumple que

$$\begin{aligned} E[SS_n(x,y)] &= E\left[\binom{n}{d+1}^{-1} \sum I\left(x,y \in S\left[x_{i_1}, x_{i_2}, \ldots, x_{i_{d+1}}\right]\right)\right] = \\ &= \binom{n}{d+1}^{-1} \sum E\left[I\left(x,y \in S\left[x_{i_1}, x_{i_2}, \ldots, x_{i_{d+1}}\right]\right)\right] = \\ &= \binom{n}{d+1}^{-1} \binom{n}{d+1} Pr\left[I\left(x,y \in S\left[X_1, X_2, \ldots, X_{d+1}\right]\right)\right] = \\ &= Pr\left[I\left(x,y \in S\left[X_1, X_2, \ldots, X_{d+1}\right]\right)\right] = SS(x,y;F). \end{aligned}$$

La versión muestral es consistente si su varianza converge a cero cuando el tamaño muestral aumenta. Esto es cierto ya que la similaridad muestral es un U-estadístico cuyo núcleo está acotado por 1, y por tanto, por el Lema 2.1 y para $r = 2$ y una constante K (que sólo depende de la cota del núcleo), se tiene que,

$$E\left[\left(SS_n(x,y) - E\left(SS_n(x,y)\right)\right)^2\right] \leq \frac{K}{n} \overset{n\to\infty}{\longrightarrow} 0. \blacksquare$$

Teorema 2.5 *La similaridad simplicial muestral es fuertemente consistente,*

$$SS_n(x,y) \underset{n\to\infty}{\overset{c.s.}{\longrightarrow}} SS(x,y;F) \ \textit{casi seguro}, \forall x, y \in \mathbb{R}^d.$$

Demostración. La demostración es similar a la del Teorema 2.4, ya que se hace uso del Lema 2.1, que muestra que, para $r \geq 2$, la similaridad muestral converge en media r-ésima. Se sabe que si la convergencia es suficientemente rápida, es decir, si

$$\sum_{n=1}^{\infty} E\left[\left|SS_n(x,y) - SS(x,y;F)\right|^r\right] < \infty,$$

entonces converge casi seguro a $SS(x,y;F)$. Por tanto, haciendo $r = 4$, se tiene que el U-estadístico verifica

$$\sum_{n=1}^{\infty} E\left[\left|SS_n(x,y) - SS(x,y;F)\right|^4\right] \leq \sum_{n=1}^{\infty} \frac{K}{n^2} < \infty,$$

lo que confirma la convergencia casi segura. $\blacksquare$

Lema 2.2 *Para cualquier función de distribución F en $\mathbb{R}^d$ y cualquier punto arbitrario pero fijo $y \in \mathbb{R}^d$, se tiene que*

$$\sup_{\|x\| \geq M} SS_n(x,y) \overset{c.s.}{\longrightarrow} 0, \ \textit{cuando } M \to \infty.$$

Demostración. Las similaridad simplicial verifica dicha convergencia ya que, como $SS\left(x,y;F\right) \leq PS\left(x;F\right)$, se tiene que

$$\sup_{\|x\|\geq M} SS_n\left(x,y\right) \leq \sup_{\|x\|\geq M} PS_n\left(x,y\right) \xrightarrow{c.s.} 0, \text{ cuando } M \to \infty,$$

por lo que la convergencia de la similaridad se cumple. ■

Lema 2.3 *Sea F es una función de distribución absolutamente continua, dado el punto $y \in \mathbb{R}^d$ arbitrario pero fijo, entonces para todo $c > 0$, se tiene que*

$$\sup_{\{x_1,x_2 \in Bola(y,c):\|x_1-x_2\|<\varepsilon\}} \left|SS_n\left(x_1,y\right) - SS_n\left(x_2,y\right)\right| \underset{\varepsilon\to 0, n\to\infty}{\longrightarrow} \gamma\left(\varepsilon\right) + R_n,$$

donde $Bola\left(y,c\right) = \left\{x \in \mathbb{R}^d : \|x-y\| \leq c\right\}$, $\gamma\left(\varepsilon\right)$ es determinista y tiende a cero y R_n converge casi seguro a cero.

Demostración. En primer lugar, se introducen los términos poblacionales y se emplea la desigualdad triangular para descomponer el valor absoluto en tres sumandos:

$$\begin{aligned}
& \left|SS_n\left(x_1,y\right) - SS_n\left(x_2,y\right)\right| \\
= \; & \left|SS_n\left(x_1,y\right) - SS\left(x_1,y;F\right) + SS\left(x_1,y;F\right) - SS_n\left(x_2,y\right)\right. \\
& \left. + SS\left(x_2,y;F\right) - SS\left(x_2,y;F\right)\right| \\
\leq \; & \left|SS_n\left(x_1,y\right) - SS\left(x_1,y;F\right)\right| + \left|SS_n\left(x_2,y\right) - SS\left(x_2,y;F\right)\right| \\
& + \left|SS\left(x_1,y;F\right) - SS\left(x_2,y;F\right)\right|
\end{aligned}$$

y, debido a que el supremo de esta suma es inferior a la suma de los supremos, el supremo original queda acotado por la suma de los tres supremos sobre los que se trabaja de forma separada.

$$\begin{aligned}
& \sup_{\{x_1,x_2 \in Bola(y,c):\|x_1-x_2\|<\varepsilon\}} \left|SS_n\left(x_1,y\right) - SS_n\left(x_2,y\right)\right| \\
\leq \; & \sup_{\{x_1,x_2 \in Bola(y,c):\|x_1-x_2\|<\varepsilon\}} \left|SS_n\left(x_1,y\right) - SS\left(x_1,y;F\right)\right| \\
& + \sup_{\{x_1,x_2 \in Bola(y,c):\|x_1-x_2\|<\varepsilon\}} \left|SS_n\left(x_2,y\right) - SS\left(x_2,y;F\right)\right| \\
& + \sup_{\{x_1,x_2 \in Bola(y,c):\|x_1-x_2\|<\varepsilon\}} \left|SS\left(x_1,y;F\right) - SS\left(x_2,y;F\right)\right|.
\end{aligned}$$

Por un lado, se tiene que $\gamma(\epsilon) = \sup_{\{x_1,x_2 \in Bola(y,c):\|x_1-x_2\|<\varepsilon\}} |SS(x_1, y; F) - SS(x_2, y; F)|$ tiende a cero cuando ε tiende a cero, debido a la que la similaridad es continua y el supremo se toma dentro del conjunto $Bola(y, c)$ que es cerrado y acotado. Por otro lado, se tiene que

$$\sup_{\{x_1,x_2 \in Bola(y,c):\|x_1-x_2\|<\varepsilon\}} |SS_n(x_2, y) - SS(x_2, y; F)|$$

es igual a

$$\sup_{A_{x_1,x_2}(y,c,\varepsilon)} |Pr_{F_n}(A_{x_1,x_2}(y, c, \varepsilon)) - Pr_F(A_{x_1,x_2}(y, c, \varepsilon))|,$$

donde $A_{x_1,x_2}(y, c, \varepsilon)$ es el conjunto de todos los símplices que contienen a los puntos x_1 e y, para x_1 tal que $\{x_1 \in Bola(y, c) : \|x_1 - x_2\| < \varepsilon\}$. Y, como los símplices del conjunto $A_{x_1,x_2}(y, c, \varepsilon)$ están contenidos en el conjunto de los que contienen a x_1 y a y $(A_{x_1,y})$, se tiene que

$$\sup_{A_{x_1,x_2}(y,c,\varepsilon)} |Pr_{F_n}(A_{x_1,x_2}(y, c, \varepsilon)) - Pr_F(A_{x_1,x_2}(y, c, \varepsilon))|$$

es igual a

$$\sup_{A_{x_1,x_2}} |Pr_{F_n}(A_{x_1,x_2}) - Pr_F(A_{x_1,x_2})|$$

lo que, unido al resultado de que la clase de todos los conjuntos medibles Borel convexos en $\mathbb{R}^d$, forman una clase Glivenko-Cantelli si F es una densidad con respecto a la medida de Lebesgue, es decir, que

$$\sup_{A \in C} |F_n(A) - F(A)| \xrightarrow{c.s.} 0,$$

donde C es el conjunto de todos los conjuntos medibles Borel convexos, da como resultado que el supremo converge casi seguro a cero. Para el punto x_2 se emplea el mismo razonamiento. ■

Teorema 2.6 *Sea F una función de distribución absolutamente continua. Entonces la similaridad $SS(x, y; F)$ es uniformemente consistente:*

$$\sup_{x \in \mathbb{R}^d} |SS_n(x, y) - SS(x, y; F)| \xrightarrow[n\to\infty]{c.s.} 0.$$

Demostración. La demostración es análoga a la del Teorema 5 en Liu (1990). Para x tal que $\|x\|$ es suficientemente grande, la propiedad (iv) de la Proposición 2.4 y el

Lema 2.2 aseguran que $|SS_n(x,y) - SS(x,y;F)| \xrightarrow[n\to\infty]{c.s.} 0$. Entonces, la afirmación ha de probarse para puntos *pequeños*, es decir, dado $M > 0$, se demostrará que

$$\sup_{x\in Q(y,M)} |SS_n(x,y) - SS(x,y;F)| \xrightarrow[n\to\infty]{c.s.} 0,$$

donde $Q(y,M)$ es un hipercubo de centro y y de longitud de lados M.
Dividiendo cada lado del hipercubo en N trozos, el hipercubo queda dividido en N^d subhipercubos. Haciendo N arbitrariamente grande y aplicando el Lema 2.3 y el Teorema 2.2, se concluye que basta con probar

$$\max_{x\in C(y,M)} |SS_n(x,y) - SS(x,y;F)| \xrightarrow[n\to\infty]{c.s.} 0,$$

donde $C(y,M)$ es el conjunto de todas las esquinas de los hipercubos.
Aplicando el Lema 2.1 con $r = 4$ y $c = 1$, se obtiene que

$$\begin{aligned}
&Pr\left(\max_{x\in C(y,M)} |SS_n(x,y) - SS(x,y;F)| > \varepsilon\right) \\
&\leq N^d \max_{x\in C(y,M)} Pr\left(|SS_n(x,y) - SS(x,y;F)| > \varepsilon\right) \\
&= N^d \max_{x\in C(y,M)} Pr\left(|SS_n(x,y) - SS(x,y;F)|^4 > \varepsilon^4\right) \\
&\leq N^d \max_{x\in C(y,M)} E\left[\varepsilon^{-4}\left(|SS_n(x,y) - SS(x,y;F)|^4\right)\right] = O\left(n^{-2}\right),
\end{aligned}$$

y empleando el lema de Borel-Cantelli se tiene que la suma sobre n de

$$Pr\left(\max_{x\in C(y,M)} |SS_n(x,y) - SS(x,y;F)| > \varepsilon\right)$$

es finita, por lo que la probabilidad del límite superior de este evento es igual a cero, concluyendo que la similaridad simplicial es fuertemente consistente. ■

A continuación se enuncian y prueban los resultados anteriores aplicados a la similaridad por bandas.

Teorema 2.7 *Dado $y \in \mathbb{R}^d$, sea F una función de distribución d-dimensional absolutamente continua entonces $SB(\cdot, y; F)$ es continua.*

Demostración. La demostración es análoga a la del Teorema 2.2. La similaridad es continua si cada sumando $SB^{b)}$ lo es. Dado $b \leq B$, la expresión

$$E\left[\prod_{k=1}^{d} I\left\{\min_{i\in\{1,2,\ldots,b\}} X_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{1,2,\ldots,b\}} X_i^{k)}\right\}\right],$$

puede reescribirse como $Pr\left(x, y \in R\left(X_1, X_2, \ldots, X_b\right)\right)$, donde $R\left(X_1, X_2, \ldots, X_b\right)$ es el menor hipercubo que contiene a los puntos $X_1, X_2, \ldots, X_b$. Entonces, definiendo los mismos sucesos que en la demostración del Teorema 2.2,

$$A_x = \left\{X_1, X_2, \ldots, X_b : x \in R\left[X_1, X_2, \ldots, X_b\right]\right\},$$

$$A_{x_n} = \left\{X_1, X_2, \ldots, X_b : x_n \in R\left(X_1, X_2, \ldots, X_b\right)\right\} \text{ y}$$

$$A_y = \left\{X_1, X_2, \ldots, X_b : y \in R\left(X_1, X_2, \ldots, X_b\right)\right\},$$

se prueba la continuidad con una sucesión convergente.

Sea $x_n \in \mathbb{R}^d$ una sucesión que converge a x, y dado un punto arbitrario pero fijo $y \in \mathbb{R}^d$ se prueba que $Pr\left(x_n, y \in R\left(X_1, X_2, \ldots, X_b\right)\right) \rightarrow Pr\left(x, y \in R\left(X_1, X_2, \ldots, X_b\right)\right)$, es decir, que $\lim_{n\to\infty} \left|Pr\left(x_n, y \in R\left(X_1, X_2, \ldots, X_b\right)\right) - Pr\left(x, y \in R\left(X_1, X_2, \ldots, X_b\right)\right)\right| = 0$. Entonces, trabajando con los conjuntos anteriores y sus complementarios se tiene por un lado que

$$A_y \cap A_{x_n} \subset \left(A_y \cap A_x\right) \cup \left(A_y \cap \overline{A_x} \cap A_{x_n}\right) \subset \left(A_y \cap A_x\right) \cup \left(\overline{A_x} \cap A_{x_n}\right),$$

y por otro que

$$A_y \cap A_x \subset \left(A_y \cap A_{x_n}\right) \cup \left(A_y \cap \overline{A_{x_n}} \cap A_x\right) \subset \left(A_y \cap A_{x_n}\right) \cup \left(\overline{A_{x_n}} \cap A_x\right).$$

En términos de probabilidades se tiene las siguientes desigualdades

$$-Pr\left(\overline{A_{x_n}} \cap A_x\right) \leq Pr\left(A_y \cap A_{x_n}\right) - Pr\left(A_y \cap A_x\right) \leq Pr\left(\overline{A_x} \cap A_{x_n}\right)$$

que combinadas, dan lugar a

$$\left|Pr\left(A_y \cap A_{x_n}\right) - Pr\left(A_y \cap A_x\right)\right| \leq Pr\left(\overline{A_x} \cap A_{x_n}\right) + Pr\left(\overline{A_{x_n}} \cap A_x\right).$$

Los sucesos $\overline{A_x} \cap A_{x_n}$ y $\overline{A_{x_n}} \cap A_x$ están incluidos en el suceso $B_{x,x_n} = \{X_1, X_2, \ldots, X_b : \exists i \in \{1, 2, \ldots, b\}, \exists k \in \{1, 2, \ldots, d\}$ tales que $\min\left(x^{k)}, x_n^{k)}\right) \leq X_i^{k)} \leq \max\left(x^{k)}, x_n^{k)}\right)\}$.

Su límite, cuando n tiende a infinito, es $B_x = \left\{ X_1, X_2, \ldots, X_b : \exists i \in \{1, 2, \ldots, b\}, \exists k \in \{1, 2, \ldots, d\} \text{ tal que } x^{k)} = X_i^{k)} \right\}$. Por último, la probabilidad de B_x es cero debido a que la función de distribución es continua. ■

Teorema 2.8 *Sea $D \subseteq \mathbb{R}^d$ un conjunto abierto y $F : D \longrightarrow \mathbb{R}^+$ una función de distribución continua. Entonces la similaridad por bandas verifica*

$$SB(x, y; F, B) = SB(y, y; F, B) \text{ sí y sólo sí } x = y.$$

Demostración. Demostración análoga a la del Teorema 2.3. Si $x = y$ la implicación hacia la izquierda se cumple. Por lo tanto es suficiente probar que $SB(x, y; F, B) = SB(y, y; F, B)$ implica que $x = y$. La prueba se realiza mediante la implicación negativa inversa, es decir, si $x \neq y$ entonces $SB(x, y; F, B) < SB(y, y; F, B)$ (ya que nunca puede ser mayor). Se prueba que el conjunto de tuplas del tipo $X_1, X_2, \ldots, X_b$ que forman bandas que contienen al punto y y no al punto x, tiene una probabilidad positiva. La similaridad por bandas está definida como

$$SB(x, y; F, B) = \sum_{b=2}^{B} E\left[\prod_{k=1}^{d} I\left\{ \min_{i \in \{1,2,\ldots,b\}} X_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i \in \{1,2,\ldots,b\}} X_i^{k)} \right\} \right],$$

donde cada sumando puede escribirse como $Pr(x, y \in R(X_1, X_2, \ldots, X_b))$, donde $R(X_1, X_2, \ldots, X_b)$ es el menor hipercubo que contiene a los puntos $X_1, X_2, \ldots, X_b$. Debido a que se tiene que $Pr(x, y \in R(X_1, \quad X_2, \ldots, X_b)) \leq Pr(y \in R(X_1, X_2, \ldots, X_b))$, entonces es suficiente probar que la desigualdad es estricta para cualquier $b \in \{2, \ldots, B\}$. Por simplicidad, y sin pérdida de generalidad, se toma $d = 2$ y $b = 2$. La prueba se realiza a través de gráficos que muestran las regiones para la la construcción de bandas que cumplan los requisitos, es decir, los puntos X_1 y X_2 tales que la región $R(X_1, X_2)$ contenga a y pero no a x. Se conclye que el conjunto de dichas bandas tiene probabilidad positiva. La continuidad de la función F y el hecho de que el conjunto D es abierto son las propiedades que aseguran que dicha probabilidad es no nula. Como puede verse en la Figura 2.27, el punto X_1 puede ser cualquiera que se encuentre en la región definida por la diferencia del conjunto D menos el cuadrante con vértice x que contiene la semirrecta $\{x + \alpha(x - y) : \alpha > 0\}$. El conjunto diferencia resultante tiene probabilidad positiva. En

el caso de que los puntos para los que se calcula la similaridad tengan el mismo valor en alguna variable la región para la elección del primer punto de la banda se realiza en el semiespacio que no contiene a dicha semirrecta. Fijado un punto X_1 perteneciente a

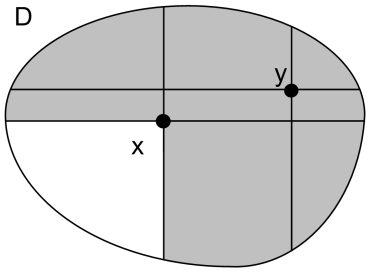


Figura 2.27: *Región para la elección de* X_1.

dicho conjunto, la región S_{X_1} formada por los puntos que producen bandas que contienen a y y no a x, queda definida como la intersección entre el cuadrante con origen en y que contiene a la semirrecta $\{y + \alpha(y - X_1) : \alpha > 0\}$ y el cuadrante con origen en x que contiene a la semirrecta $\{x + \alpha(y - x) : \alpha > 0\}$. Dicha región, tal como se observa en la Figura 2.28, tiene probabilidad no nula, por lo que la función es continua. ■

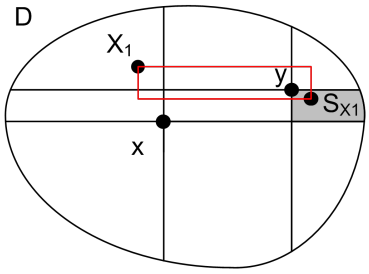


Figura 2.28: *Región* S_{X_1} *para la elección de* X_2.

Teorema 2.9 *La similaridad por bandas es insesgada y consistente*

$$SB_n\left(x, y; B\right) \underset{n\to\infty}{\overset{p}{\to}} SB\left(x, y; F, B\right), \ \forall x, y \in \mathbb{R}^d.$$

Demostración. La esperanza de $SB_n\left(x, y; B\right)$ es la suma de las esperanzas de sus

sumandos, así pues, para $b = 1, 2, \ldots, B$ se tiene que $E\left[SB_n^{b)}(x,y)\right]$

$$
\begin{aligned}
&E\left[\frac{1}{\binom{n}{b}} \sum_{1\leq i_1<\ldots<i_b\leq n} \prod_{k=1}^{d} I\left\{\min_{i\in(i_1,\ldots,i_b)} x_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in(i_1,\ldots,i_b)} x_i^{k)}\right\}\right] \\
&= \binom{n}{b}^{-1} \sum_{1\leq i_1<\ldots<i_b\leq n} E\left[I\left\{\min_{i\in(i_1,\ldots,i_b)} x_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in(i_1,\ldots,i_b)} x_i^{k)}, \forall k \in \{1,2,\ldots,d\}\right\}\right] \\
&= \binom{n}{b}^{-1} \sum_{1\leq i_1<\ldots<i_b\leq n} Pr\left(x, y \in R\left(X_1, X_2, \ldots, X_b\right)\right) \\
&= Pr\left(x, y \in R\left(X_1, X_2, \ldots, X_b\right)\right)
\end{aligned}
$$

y su suma es $SB\left(x, y; F, B\right)$. Para probar la consistencia de la similaridad se hace uso del Lema 2.1 para $r = 2$, debido a que todos los sumandos de la similaridad son U-estadísticos de función núcleo acotada por 1. Aplicando el resultado se tiene que la varianza de cada sumando verifica que

$$E\left[\left(SB_n^{b)}(x,y) - E\left[SB_n^{b)}(x,y)\right]\right)^2\right] \leq \frac{K}{n} \underset{n\to\infty}{\longrightarrow} 0,$$

para una constante K que sólo depende de la cota de la función núcleo del U-estadístico, donde $SB^{b)}$ denota el sumando asociado a b bandas. Cada $SB^{b)}$ converge en probabilidad, y, por lo tanto, su suma también. ■

Teorema 2.10 *La similaridad por bandas es fuertemente consistente*

$$SB_n\left(x, y; B\right) \underset{n\to\infty}{\overset{c.s.}{\to}} SB\left(x, y; F, B\right), \ \forall x, y \in \mathbb{R}^d.$$

Demostración. Por medio del Lema 2.1, tomando $r = 4$, se tiene para cada sumando que

$$E\left[\left(SB_n^{b)}(x,y) - E\left[SB_n^{b)}(x,y)\right]\right)^4\right] \leq \frac{K}{n^2} \underset{n\to\infty}{\longrightarrow} 0,$$

y, por lo tanto,

$$\sum_{n=1}^{\infty} E\left[\left(SB_n^{b)}(x,y) - E\left[SB_n^{b)}(x,y)\right]\right)^4\right] < \infty.$$

Cada sumando converge casi seguro a su valor poblacional debido a que la convergencia es suficientemente rápida. Finalmente, como la suma de variables que convergen casi seguro es también convergente casi seguro se tiene la convergencia de la similaridad. ■

Lema 2.4 *Para cualquier F en $\mathbb{R}^d$ y un punto arbitrario pero fijo $y \in \mathbb{R}^d$, se tiene que*

$$\sup_{\|x\| \geq M} SB_n(x, y; B) \xrightarrow{c.s.} 0, \text{ cuando } M \to \infty.$$

Demostración. Se sabe que la profundidad por bandas verifica esta convergencia. Por lo tanto, debido a que $SB(x, y; F, B) \leq PB(x; F, B)$, se tiene que

$$\sup_{\|x\| \geq M} SB_n(x, y; B) \leq \sup_{\|x\| \geq M} PB_n(x; B) \xrightarrow{c.s.} 0, \text{ cuando } M \to \infty,$$

quedando probada la convergencia para SB_n. ■

Lema 2.5 *Sea F una función de distribución d-dimensional absolutamente continua, para cada $y \in \mathbb{R}^d$ se tiene que $\forall c > 0$,*

$$\sup_{\{x_1, x_2 \in B(y,c): \|x_1 - x_2\| < \varepsilon\}} |SB_n(x_1, y; B) - SB_n(x_2, y; B)| \xrightarrow[\varepsilon \to 0, n \to \infty]{c.s.} 0,$$

donde $B(y, c) = \{x : \|x - y\| \leq c\}$.

Demostración. El razonamiento para la demostración es el mismo que el de la demostración del Lema 2.3, por lo que se omite. ■

Teorema 2.11 *Sea F una función de distribución absolutamente continua. Entonces $SB_n(x, y; B)$ es uniformemente consistente,*

$$\sup_{x \in \mathbb{R}^d} |SB_n(x, y; B) - SB(x, y; F, B)| \xrightarrow[n \to \infty]{c.s.} 0.$$

Demostración. La demostración es análoga a la del teorema 5 de Liu (1990). Para x tal que $\|x\|$ es suficientemente grande gracias a la propiedad (iv) de la Proposición 2.5 y al Lema 2.4 se tiene que $|SB(x, y; B) - SB(x, y; F, B)| \xrightarrow[n \to \infty]{c.s.} 0$. Se prueba ahora para puntos de menor magnitud, es decir, dada $M > 0$, se prueba que

$$\sup_{x \in Q(y,M)} |SB_n(x, y; B) - SB(x, y; F, B)| \xrightarrow[n \to \infty]{c.s.} 0 \text{ cuando } n \to \infty,$$

donde $Q(y, M)$ es un hipercubo de centro y y de lado M.

Dividiendo cada arista en N trozos, el hipercubo queda dividido en N^d subhipercubos.

Haciendo N arbitrariamente grande y debido a los resultados del Lema 2.3, del Teorema 2.2 y del lema de Borel-Cantelli, basta probar que

$$\underset{x\in C(y,M)}{\text{máx}} \left|SB_n\left(x,y;B\right) - SB\left(x,y;F,B\right)\right| \xrightarrow[n\to\infty]{c.s.} 0 \text{ } cuando \text{ } n\to\infty,$$

donde $C\left(y,M\right)$ es el conjunto de todos los vértices de los subhipercubos.

Aplicando el Lema 2.1 con $r=4$ y $c=1$, resulta que

$$\begin{aligned}
& Pr\left(\underset{x\in C(y,M)}{\text{máx}} \left|SB_n\left(x,y;B\right) - SB\left(x,y;F,B\right)\right| > \varepsilon\right) \\
\leq\ & N^d \underset{x\in C(y,M)}{\text{máx}} Pr\left(\left|SB_n\left(x,y;B\right) - SB\left(x,y;F,B\right)\right| > \varepsilon\right) \\
=\ & N^d \underset{x\in C(y,M)}{\text{máx}} Pr\left(\left|SB_n\left(x,y;B\right) - SB\left(x,y;F,B\right)\right|^4 > \varepsilon^4\right) \\
\leq\ & N^d \underset{x\in C(y,M)}{\text{máx}} E\left[\varepsilon^{-4}\left(\left|SB_n\left(x,y;B\right) - SB\left(x,y;F,B\right)\right|^4\right)\right] = O\left(n^{-2}\right),
\end{aligned}$$

y, por el lema de Borel-Cantelli, la suma sobre n de $Pr\left(\text{máx}_{x\in C(y,M)} \left|SB_n\left(x,y;B\right)\right.\right.$ $\left.\left. -SB\left(x,y;F,B\right)\right| > \varepsilon\right)$ es finita. Por tanto la probabilidad del límite superior de este suceso es igual a cero, por lo que la similaridad por bandas es fuertemente consistente. ■

A continuación se presentan resultados para la similaridad por bandas modificada. Para las demostraciones de estos resultados se hace uso de la notación introducida en la Observación 2.2.

Teorema 2.12 *Dado $y\in\mathbb{R}^d$, sea F es una función de distribución d-dimensional absolutamente continua entonces $SBM\left(\cdot,y;F,B\right)$ es continua.*

Demostración. Se prueba que cada sumando $SBM^{b,k)}\left(x^{k)},y^{k)};F\right)$, con $b=2,\ldots,B$ y $k=1,2,\ldots,d$, es continuo y, debido a que SBM es una función lineal de estos sumandos, entonces será también continua. Para ver si cada sumando es continuo, se analizan en términos de probabilidades: $SBM^{b,k)}\left(x^{k)},y^{k)};F\right)$ es la probabilidad de que el intervalo construido a partir del mínimo y el máximo de las k-ésimas coordenadas de una muestra aleatoria simple de tamaño b, contenga a la k-ésima coordenada de x e y, es decir,

$$Pr\left[\text{mín}\left(X_1^{k)},X_2^{k)},\ldots,X_b^{k)}\right) \leq y^{k)},x^{k)} \leq \text{máx}\left(X_1^{k)},X_2^{k)},\ldots,X_b^{k)}\right)\right].$$

Sea $x_n\in\mathbb{R}^d$ una sucesión tal que $x_n\to x$. Se comprueba que la diferencia

$$\left|SBM^{b,k)}\left(x_n^{k)},y^{k)};F\right) - SBM^{b,k)}\left(x^{k)},y^{k)};F\right)\right|$$

tiende a cero. A esta diferencia contribuyen las observaciones cuyos intervalos contienen a $y^{k)}$ y, o sólo a $x^{k)}$ o sólo a $x_n^{k)}$. Pero, debido a que el $\lim_{n\to\infty} x_n^{k)} = x^{k)}$, los límites de los eventos que contribuyen a la diferencia son $\min\left(X_1^{k)}, X_2^{k)}, \ldots, X_b^{k)}\right) = x^{k)}$ y $\max\left(X_1^{k)}, X_2^{k)}, \ldots, X_b^{k)}\right) = x^{k)}$ los cuales tienen probabilidad nula debido a la continuidad de F. ■

Teorema 2.13 *Sea $D \subseteq \mathbb{R}^d$ un abierto y $F : D \longrightarrow \mathbb{R}^+$ una función de distribución continua. Entonces la similaridad por bandas modificada verifica*

$$SBM\,(x, y; F, B) = SBM\,(y, y; F, B) \text{ si, y sólo si, } x = y.$$

Demostración. Si $x = y$, se cumple que $SBM\,(x, y; F, B) = SBM\,(y, y; F, B)$. Para la implicación opuesta se prueba que $x \neq y$ implica la desigualdad del lado izquierdo. Debido a que $SBM^{b,k)}\left(x^{k)}, y^{k)}; F\right) \leq SBM^{b,k)}\left(y^{k)}, {}^{k)}; F\right)$ para $b = 2, 3, \ldots, B$ y $k = 1, 2, \ldots, d$, entonces es suficiente probar la desigualdad estricta para algún valor de b y k. Por simplicidad se toman $d = 2$, $b = 2$, para $k = 1$ y $k = 2$ simultáneamente.
Dados x e y la región para la elección de uno de los dos puntos que formen la banda se representa en la Figura 2.29. Para cualesquiera $x, y \in \mathbb{R}^d$ esta región es el cuadrante con origen en x que contiene al punto y. Este cuadrante siempre tendrá probabilidad no nula. Para cualquier punto en el cuadrante se tendrá una región no vacía para seleccionar el

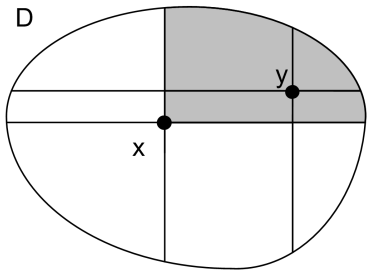


Figura 2.29: *Región S_{X_1} para la elección de X_2.*

otro punto de la banda, como puede verse en la Figura 2.30. ■

Teorema 2.14 *La similaridad por bandas modificada es insesgada y consistente,*

$$SBM_n\,(x, y; B) \xrightarrow[n\to\infty]{p} SBM\,(x, y; F, B)\,, \ \forall x, y \in \mathbb{R}^d.$$

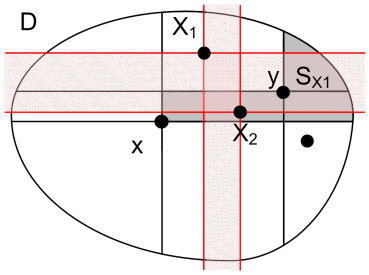


Figura 2.30: *Región S_{X_1} para la elección de X_2.*

Demostración. $SBM_n\left(x, y; B\right)$ será insesgada si cada sumando lo es. Se tiene que $E\left[SBM_n^{b,k)}\left(x^{k)}, y^{k)}\right)\right]$ es igual a

$$E\left[\binom{n}{b}^{-1}\sum_{(i_1,i_2,\ldots,i_b)\in J_b} I\left\{\min_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)}\right\}\right]$$
$$= \binom{n}{b}^{-1}\sum_{(i_1,i_2,\ldots,i_b)\in J_b} Pr\left(\min_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)}\right)$$
$$= Pr\left(\min_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)} \leq x^{k)}, y^{k)} \leq \max_{i\in\{i_1,i_2,\ldots,i_b\}} x_i^{k)}\right) = SBM^{b,k)}\left(x^{k)}, y^{k)}; F\right),$$

donde J_b es el conjunto de todas las combinaciones de n puntos tomados de b en b y, por lo tanto,

$$E\left[\frac{1}{d}\sum_{b=2}^{B}\sum_{k=1}^{d} SBM_n^{b,k)}\left(x^{k)}, y^{k)}\right)\right] = \frac{1}{d}\sum_{b=2}^{B}\sum_{k=1}^{d} E\left[SBM_n^{b,k)}\left(x^{k)}, y^{k)}\right)\right] = SBM\left(x, y; F, B\right).$$

Aplicando el Lema 2.1 con $r = 2$ a cada sumando, se tiene que la varianza del sumando $SBM_n^{b,k)}\left(x^{k)}, y^{k)}\right)$ tiende a cero y por lo tanto $SBM_n^{b,k)}\left(x^{k)}, y^{k)}\right) \xrightarrow{p} SBM^{b,k)}\left(x^{k)}, y^{k)}; F\right)$ lo que implica que $SBM_n\left(x, y; B\right) \xrightarrow{p} SBM\left(x, y; F, B\right)$. ■

Teorema 2.15 *La similaridad por bandas modificada es fuertemente consistente,*

$$SBM_n\left(x, y; B\right) \underset{n\to\infty}{\overset{c.s.}{\to}} SBM\left(x, y; F, B\right), \ \forall x, y \in \mathbb{R}^d.$$

Demostración. La demostración es análoga a las de los teoremas 2.5 y 2.10. Se prueba que cada sumando verifica la convergencia. Aplicando el Lema 2.1 con $r = 4$ y $m = b$, la varianza de $SBM_n^{b,k)}\left(x^{k)}, y^{k)}\right)$ para $k = 1, 2, \ldots, d$ decrece suficientemente

rápido como para verificar $\sum_{n=1}^{\infty} E\left[\left(SBM_n^{b,k)}\left(x^{k)}, y^{k)}\right) - E\left[SBM_n^{b,k)}\left(x^{k)}, y^{k)}\right)\right]\right)^4\right] < \infty$. ■

Como se ha comentado en la sección 2.4.1, la similaridad por bandas modificada no se desvanece en el infinito, por lo que la afirmación $\sup_{\|x\|\geq M} SBM\left(x, y; F_n, B\right) \xrightarrow{c.s.} 0$, cuando $M \to \infty$ no es cierta.

Para finalizar con las propiedades de las funciones propuestas es conveniente estudiarlas como funciones de similaridad. Una función de similaridad $S\left(\cdot,\cdot\right)$ tiene que verificar las tres propiedades siguientes para cualquier par de puntos a y b:

1. $S\left(a,a\right) \geq S\left(a,b\right)$
2. $S\left(a,b\right) \geq 0$
3. $S\left(a,a\right) = S\left(a,b\right) \Leftrightarrow a = b$

Por definición, y como se ha comprobado previamente, las dos primeras propiedades se cumplen para todas las similaridades propuestas. Para las similaridades de Mahalanobis, Oja y por proyecciones, la tercera propiedas siempre se satisface. Para el resto, se cumplirán siempre que la función de distribución sea absolutamente continua.

2.5. Aplicación de las similaridades en el análisis de conglomerados jerárquico

El análisis de conglomerados se enmarca dentro de los métodos de clasificación no supervisados, cuyo objetivo es clasificar y agrupar las observaciones cuando no se conoce la pertenencia de éstas a los posibles grupos. Es una técnica exploratoria, que trata de describir cómo se encuentran las observaciones en el espacio. Algunos de los métodos de análisis de conglomerados necesitan información a priori acerca del número de grupos en que se divide el conjunto, antes de llevar a cabo la agrupación de las observaciones. Los métodos más empleados se dividen en dos grupos: jerárquicos y de particionado.

Los métodos de particionado, realizan una partición en K grupos, donde K es generalmente definido por el usuario, aunque su elección se puede automatizar. Se considera

que un agupamiento es partición si cada uno de los grupos contiene al menos una observación y cada observación pertenece a un único grupo. Con estas condiciones se tiene que, como mucho, podrá haber tantos grupos como observaciones. El objetivo es llevar a cabo una partición adecuada, es decir, una partición en la que los grupos sean entre sí lo más heterogeneos posible e internamente presenten una elevada homogeneidad. Ejemplos de este tipo de métodos son, por ejemplo, k-medias Hartigan (1975), PAM (Partitioning Around Medoids, Kaufman y Rousseeuw (1987)), CLARA (Clustering Large Applications, Kaufman y Rousseeuw (1986)), FANNY (Fuzzy Analysis, Kaufman y Rousseeuw (1990)) y, basado en la profundidad L_1, el propuesto en Jornsten (2004).

El otro grupo de técnicas de análisis de conglomerados es el de los métodos jerárquicos. Éstos, a diferencia de los de particionado, no ofrecen una única división del conjunto de observaciones, ya que el método es iterativo y en cada paso se obtiene un agrupamiento. Existen dos tipos de conglomerados jerárquicos, los aglomerativos y los divisivos. En el primer tipo se comienza con tantos grupos como observaciones y se van aglomerando observaciones y subgrupos de forma iterativa hasta tener un único grupo compuesto por todas las observaciones. Los divisivos actúan en sentido contrario: comienzan con un primer grupo formado por toda la muestra y, de forma iterativa, divide todos los subgrupos hasta obtener tantos grupos como elementos de la muestra. Todos los pasos de aglomeración o división que se llevan a cabo son representados por medio de árboles conocidos como dendrogramas. A partir de estos árboles es posible obtener las agrupaciones para determinados valores del número de grupos.

Existen varios criterios para la determinación de las distancias entre grupos. Los más importantes son:

1. Encadenamiento simple o vecino más próximo: Se toma la distancia entre dos grupos como la menor distancia entre elementos de ambos grupos.

2. Encadenamiento completo o vecino más alejado: Se toma la distancia entre dos grupos como la mayor distancia entre elementos de ambos grupos.

3. Media de grupos: Se toma la distancia entre dos grupos como la media de las

distancias entre elementos de ambos grupos.

4. Método de Ward: Se parte con tantos grupos como observaciones. En cada paso se agrupan los objetos que produzcan el incremento mínimo de la suma sobre todos los grupos de la suma del cuadrado de las distancias entre elementos del grupo y su centroide.

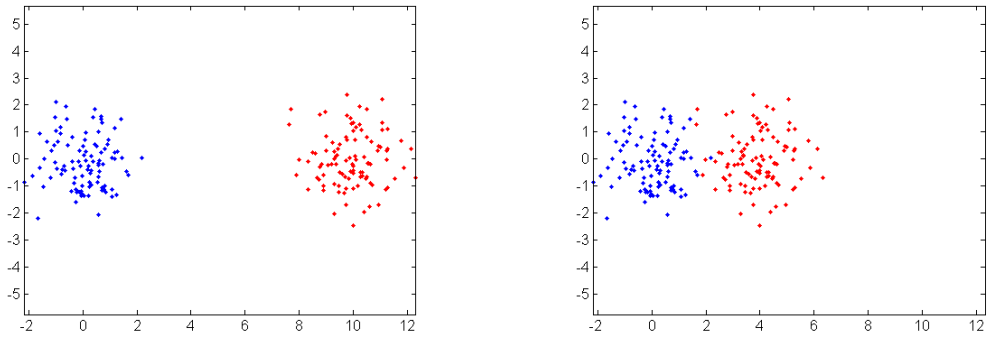
En esta sección se aplican las distintas similaridades propuestas en el análisis de conglomerados jerárquico. Se analizan diferentes conjuntos de datos en dimensión dos, simulados sobre distintas distribuciones y configuraciones de los grupos. Los resultados se comparan con los obtenidos con la distancia euclídea.

2.5.1. Cálculo de la matriz de similaridades

Las similaridades definidas en este capítulo son adecuadas para obtener las proximidades entre pares de puntos conforme a la forma de la nube de puntos o de la distribución, aunque algunas de ellas tienen, como se ha visto previamente, una mayor capacidad para adaptarse a éstas. Éste es el caso de las similaridades por bandas y por bandas modificada. Cuando se trata de describir proximidades de una forma descriptiva sobre una muestra, su comportamiento es bueno. Sin embargo, al aplicarlas en la búsqueda de grupos o conglomerados los resultados no son adecuados. Esto se debe a que no hay observaciones en los huecos que separan a los grupos, lo que hace que la similaridad muestral los ignore y estime que la similaridad entre estos puntos sea alta a pesar de que la distancia entre ellos sea elevada.

Para ilustrar este fenómeno se presenta la Figura 2.31. Ésta contiene los diagramas de dispersión de dos muestras generadas a partir de una mixtura de normales. En el primero de los diagramas se grafican las observaciones de la muestra cuyos grupos están claramente separados. En el segundo, las observaciones de la muestra cuyos grupos son tangenciales. En esta situación las similaridades por bandas y por bandas modificada fallan a la hora de identificar las disimilaridades entre objetos de un grupo y los del otro. Esto puede verse claramente en la Tabla 2.1, en la que se presentan por un lado

las similaridades muestrales medias entre observaciones de grupos diferentes y por otro, las similaridades medias entre elementos del mismo grupo. Se observa en primer lugar, que la diferencia entre las medias entre e intra grupos para ambos ejemplos es menor para la similaridad por bandas modificada. En segundo, que tanto para la similaridad por bandas como para la similaridad por bandas modificadas apenas hay diferencia en los valores medios (entre vs. entre e intra vs. intra) entre los dos ejemplos. También la similaridad por proyecciones presenta escasas diferencias. Por último para la simplicial, en terminos absolutos no muestra diferencias elevadas, su cambio sí se observa en términos relativos, en lo referente a las similaridades entre grupos.



(a) Grupos separados (b) Grupos cercanos

Figura 2.31: *Diagramas de puntos para grupos distribuidos normales próximos y distantes.*

Para solucionar este problema presente en las similaridades por bandas y por bandas modificada, se propone completar el espacio con una función de distribución continua. De este modo se introduce masa de probabilidad en los huecos, con lo que se consigue separar los grupos. En la práctica, este completado de espacio consiste en calcular las similaridades no con respecto a la muestra, sino con respecto a otra distribución. Pero, como una de las ventajas que ofrecen tanto profundidades como similaridades basadas en profundidad muestrales es que tienen en cuenta la forma de la nube de puntos, no parecere del todo adecuado cambiar totalmente la distribución empleada para el cálculo. Por tanto, se propone emplear una mixtura entre la función de distribución muestral y la función de relleno.

	Grupos alejados		Grupos próximos	
Similaridad	Entre grupos	Intra grupos	Entre grupos	Intra grupos
Mahalanobis	0.1648	0.5234	0.2003	0.4596
Proyecciones	0.2466	0.4313	0.2576	0.3973
Oja	0.3649	0.5523	0.5870	0.6966
Simplicial	0.0417	0.2117	0.0723	0.1913
Bandas	0.1558	0.2901	0.1574	0.2919
Bandas modificada	0.4176	0.5552	0.4241	0.5588

Tabla 2.1: *Media de las similaridades con respecto a la distribución muestral.*

Esta distribución de relleno no debe distorsionar en gran medida la función de distribución empírica, ya que en ese caso se podrían obtener resultados no deseados. Así pues, un requisito deseable es que los parámetros de la función de relleno se estimen por medio de la muestra, para que completen el espacio de forma coherente. Por ejemplo, es posible tomar la distribución normal multivariante con vector de medias y matriz de covarianzas muestrales o una distribución t de Student si se desea que tenga colas más pesadas. Otra opción es rellenar de forma uniforme un hipercubo (o elipse) que contenga a todas las observaciones.

En esta sección se utiliza la distribución normal multivariante como función de completado y se le da el mismo peso a ésta que a la empírica, es decir, las similaridades se calculan con respecto a mixtura $F = \alpha \cdot Normal\left(\hat{\mu}, \hat{\Sigma}\right) + (1-\alpha)\cdot F_n$, con $\alpha = 0.5$ y donde $\hat{\mu}$ y $\hat{\Sigma}$ son, respectivamente, el vector de medias y la matriz de covarianzas muestral.

La Tabla 2.2 contiene las similaridades medias de la Tabla 2.1 tomando como distribución de cálculo de las similaridades esta mixtura. Se observa cómo, para las similaridades por bandas y por bandas modificada, se produce un cambio en las similaridades entre grupos al pasar de grupos próximos a grupos alejados que, si bien no es elevado en términos absolutos (-0.0117 en SB y -0.018 en SBM), sí lo es en términos relativos (-8.6 % en SB y -4.6 % en SBM). Por otro lado se puede observar que, para la mayoría de similaridades, las medias de similaridades entre grupos disminuye tras emplear la mixtura. Para

las medias de las similaridades intra grupos se produce mayoritariamente un incremento. Esto significa que al aplicar la mixtura se consigue que los elementos de un mismo grupo sean más similares entre sí y menos con los del otro grupo.

	Grupos alejados		Grupos próximos	
Similaridad	Entre grupos	Intra grupos	Entre grupos	Intra grupos
Mahalanobis	0.1639	0.5227	0.2002	0.4597
Proyecciones	0.2427	0.4364	0.2538	0.3999
Oja	0.3661	0.5472	0.5882	0.6968
Simplicial	0.0391	0.2335	0.0654	0.1952
Bandas	0.1236	0.3401	0.1353	0.3042
Bandas modificada	0.3770	0.5991	0.3950	0.5674

Tabla 2.2: *Media de las similaridades con respecto a la mixtura.*

Por último, otra gran ventaja que se produce al introducir el relleno es que, si la función de relleno es continua en todo el espacio, las similaridades calculadas para puntos fuera de la envolvente convexa en el caso simplicial, fuera del menor hipercubo que contiene a toda la muestra en el caso de la similaridad por bandas o fuera de los intervalos de todas las coordenadas en el caso de la similaridad por bandas modificada, no son iguales a cero, esquivando de este modo uno de los mayores inconvenientes de estas tres similaridades. Como desventajas principales están que el cálculo, para la mayoría de las similaridades, presenta un incremento en el coste computacional y que sólo puede obtenerse una aproximación. Para las similaridades por bandas y por bandas modificadas no se da ninguno de estos dos problemas.

Dada una muestra de n observaciones $x_1, x_2, \ldots, x_n$ en $\mathbb{R}^d$ se obtiene la matriz de similaridades

$$S = \begin{pmatrix} S_{1,1} & S_{1,2} & \ldots & S_{1,n} \\ S_{1,2} & S_{2,2} & \ldots & S_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,1} & S_{n,2} & \ldots & S_{n,n} \end{pmatrix},$$

donde $S_{i,j}$ representa la similaridad entre el punto x_i y el punto x_j con respecto a la mixtura $\alpha \cdot Normal\left(\hat{\mu}, \hat{\Sigma}\right) + (1-\alpha) \cdot F_n$, donde $\hat{\mu}$ y $\hat{\Sigma}$ el vector de medias y la matriz de covarianzas muestrales y $\alpha \in [0,1]$.

Una vez se ha calculado la matriz de similaridades, es necesario transformarla en otra de disimilaridades, ya que el análisis de conglomerados jerárquicos se obtiene a partir de matrices de este tipo.

Debido a que las similaridades simplicial, por bandas y por bandas modificadas, el rango de posibles valores está acotado por el valor de la profundidad de los puntos que se comparan, se realiza un escalado con el que se consigue que los nuevos valores puedan tomar cualquier valor en el intervalo $[0,1]$. Con este escalado se consigue además que las similaridades entre pares de puntos distintos estén en la misma escala y sean, por tanto, comparables.

Sea D la matriz diagonal

$$D = \begin{pmatrix} S_{1,1} & 0 & \dots & 0 \\ 0 & S_{2,2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & S_{n,n} \end{pmatrix},$$

se obtiene la similaridad escalada $S^{esc} = D^{-1/2} S D^{-1/2}$.

Finalmente, dada la matriz de similaridad escalada con unos en la diagonal principal, S^{esc}, se aplica la transformación logarítmica para obtener las disimilaridades,

$$\delta = \begin{pmatrix} -log(S^{esc}_{1,1}) & -log(S^{esc}_{1,2}) & \dots & -log(S^{esc}_{1,n}) \\ -log(S^{esc}_{1,2}) & -log(S^{esc}_{2,2}) & \dots & -log(S^{esc}_{2,n}) \\ \vdots & \vdots & \ddots & \vdots \\ -log(S^{esc}_{n,1}) & -log(S^{esc}_{n,2}) & \dots & -log(S^{esc}_{n,n}) \end{pmatrix}.$$

2.5.2. Ejemplos de aplicación

En este apartado se presentan los resultados obtenidos tras aplicar el análisis de conglomerados jerárquico sobre muestras formadas por varios grupos y generadas de forma

aleatoria. Debido a que uno de los objetivos es poder comparar entre sí las similaridades propuestas, todos los conjuntos de datos son de dimensión dos, ya que, en dimensiones mayores, algunas de las similaridades no se pueden obtener en un tiempo razonable. Los resultados para las similaridades se comparan con los obtenidos tomando la distancia Euclídea. En todos los ejemplos se ha utilizado la mixtura al 50 % con la distribución normal multivariante.

En cuanto al criterio de determinación de la distancia entre grupos empleado en el análisis, la experiencia sugiere para todas las similaridades la elección del método de Ward. Para la distancia euclídea se toma tanto este criterio, como el del vecino más próximo.

En total se analizan 14 conjuntos de datos, compuestos por dos, tres o cuatro grupos. Dependiendo del ejemplo, los grupos tienen distribuciones simétricas, asimétricas e incluso con relaciones no lineales. En ningún caso se representan los dendrogramas, pues el tamaño muestral de los conjuntos de datos no permite su representación de forma adecuada, aunque se utilizan para hacer la división según el número de grupos que corresponda. Para cada similaridad y para cada criterio de agrupación en el caso de la distancia euclídea, se grafican las observaciones caracterizando las distintas agrupaciones resultantes mediante colores.

Se comienza mostrando los resultados para los ejemplos cuyos grupos tienen distribución simétrica.

2.5.2.1. Grupos con distribución simétrica

El primer grupo de conjunto de datos está compuesto por agrupaciones simétricas. Se presentan seis ejemplos de datos simulados. En los cuatro primeros, los grupos se han generado a partir de la distribución normal bivariante y en todos, salvo en uno en que hay tres, están compuestos por dos grupos. En los otros dos ejemplos, los grupos tienen forma rectangular y han sido generados a partir de vectores aleatorios con coordenadas independientes distribuidas según una uniforme. Uno de ellos contiene dos grupos y el otro cuatro. Los ejemplos son los siguientes.

Ejemplo 1: Dos grupos de tamaños $n_1 = n_2 = 100$, simulados a partir de las distribuciones $Normal\,(\mathbf{0}, \mathbf{I})$ y $Normal\,(\mu, \mathbf{I})$, donde $\mathbf{0}$ es un vector de ceros de dimensión dos, $\mathbf{I}$ es la matriz identidad de dimensión 2×2 y $\mu = (5, 5)'$.

Ejemplo 2: Dos grupos cuyas variables tienen correlaciones opuestas, de tamaños $n_1 = n_2 = 100$ y simulados a partir de las distribuciones $Normal\,(\mathbf{0}, \Sigma_1)$ y $Normal\,(\mu, \Sigma_2)$, donde $\mathbf{0}$ es un vector de ceros de dimensión dos, $\mu = (3.5,\ 3.5)'$,

$$\Sigma_1 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \text{ y } \Sigma_2 = \begin{pmatrix} 1 & \text{-}0.8 \\ \text{-}0.8 & 1 \end{pmatrix}.$$

Ejemplo 3: Dos grupos con variables de dispersión diferentes, de tamaños $n_1 = n_2 = 100$ y simulados a partir de las distribuciones $Normal\,(\mathbf{0}, \Sigma)$ y $Normal\,(\mu, \Sigma)$, donde $\mathbf{0}$ es un vector de ceros de dimensión dos, $\mu = (5, 0)'$ y

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}.$$

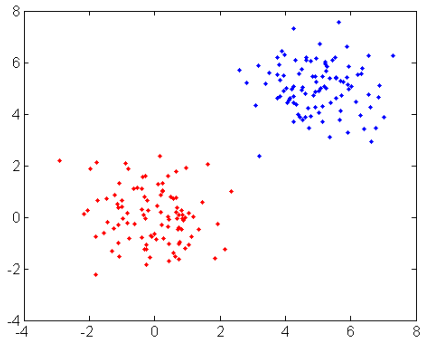
Ejemplo 4: Tres grupos con variables de dispersión diferentes, de tamaños $n_1 = n_2 = n_3 = 100$. Dos de los grupos son los del ejemplo 3. El otro ha sido simulado a partir de $Normal\,(\mu, \mathbf{I})$, donde $\mu = (10, 20)'$ e $\mathbf{I}$ es la matriz identidad de dimensión 2×2.

Ejemplo 5: Dos grupos rectangulares con coordenadas uniformes, de tamaños $n_1 = n_2 = 100$. El primer grupo tiene la primera coordenada con distribución U (-0.5, 0.5) y la segunda con U (0.5, 10.5). El segundo grupo tiene la primera coordenada distribuida según una U (0.5, 10.5) y la segunda según una U (-0.5, 0.5).

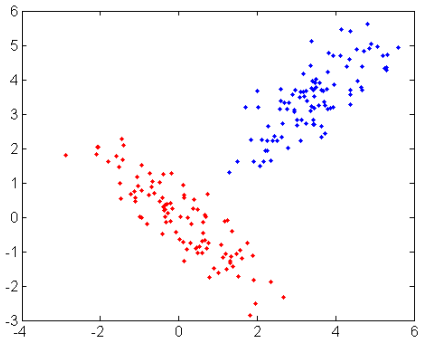
Ejemplo 6: Cuatro grupos rectangulares con coordenadas uniformes, de tamaños $n_1 = n_2 = n_3 = n_4 = 100$. Dos de estos grupos han sido generados a partir de las distribuciones del ejemplo 5. El tercer grupo tiene la primera coordenada con distribución U (-0.5, 0.5) y la segunda con U (-0.5, -10.5) y el cuarto tiene la primera coordenada distribuida según una U (-10.5, -0.5) y la segunda según una U (-0.5, 0.5).

La Figura 2.32 contiene los diagramas de puntos de estos seis ejemplos. Puede observarse que en todos los ejemplos los grupos, salvo para alguna observación, son separables.

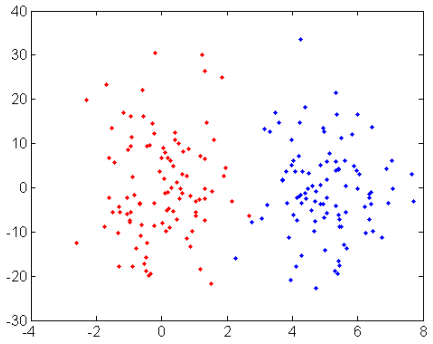
Las Figuras desde 2.33 hasta 2.38, contienen para cada similaridad y para la distancia euclídea (métodos de encadenamiento simple y de Ward) las agrupaciones resultantes de



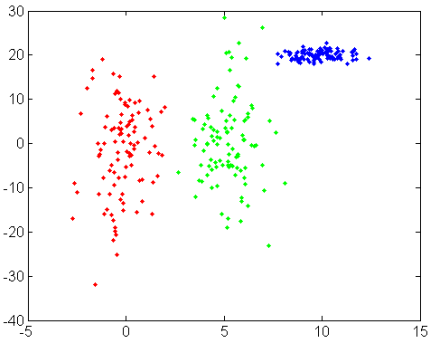
(a) Ejemplo 1



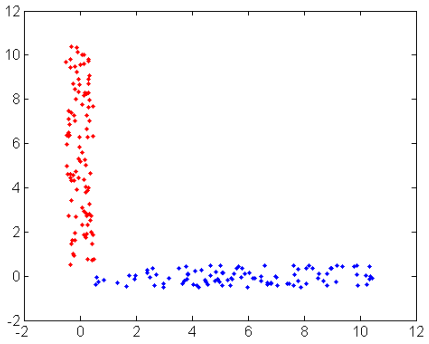
(b) Ejemplo 2



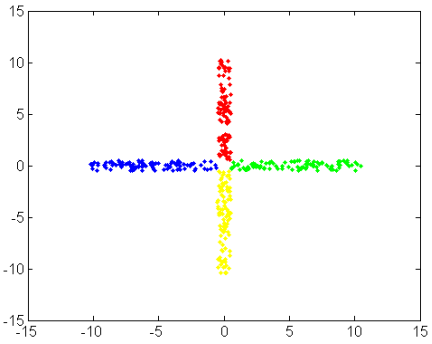
(c) Ejemplo 3



(d) Ejemplo 4



(e) Ejemplo 5



(f) Ejemplo 6

Figura 2.32: *Diagramas de dispersión de los ejemplos con grupos simétricos.*

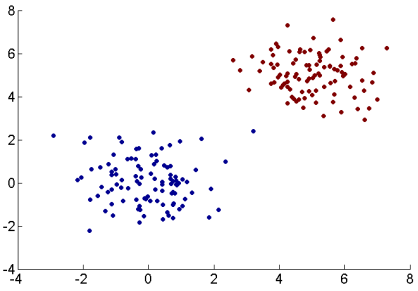
aplicar un corte al dendrograma a la altura correspondiente para dividir la muestra en el número de grupos de cada ejemplo.

Para el ejemplo más sencillo, Figura 2.33, se tiene que, salvo para la similaridad simplicial en la que se clasifica de forma incorrecta una observación del grupo de media cero, todas las demás funciones clasifican incorrectamente una observación del grupo de vector de medias $\mu = (5, 5)'$. El error es por tanto para todas las funciones de 0.5 %. Para el ejemplo número dos (Figura 2.34), el de grupos con variables correladas, todas las funciones menos la de similaridad por proyecciones que clasifica mal dos observaciones (1 %), no cometen ningún error. Si una de las variables posee una variabilidad mucho mayor que la otra (Figura 2.35), la distancia euclídea en el mejor de los casos parte ambos grupos y agrupa las mitades, obteniéndose en torno a un 50 % de error. En ese caso, las similaridades que peor se comportan son por bandas y por bandas modificada con 9.5 y 5.5 por ciento de error, respectivamente. Si se añade otro grupo de media distinta y con sus dos coordenadas de igual varianza, Figura 2.36, la distancia euclídea con el método de encadenamiento simple llega a obtener un error del 66 % (con Ward del 29.7 %). En este caso, la similaridad que peor agrupamiento ofrece es la de Oja, con un 16 % de error, seguida de las de Mahalanobis y por proyecciones con un 10 %.

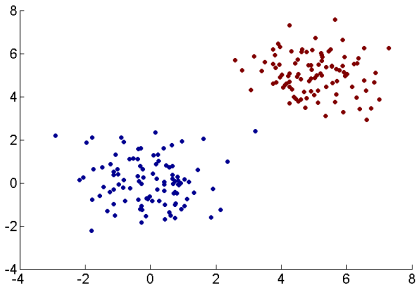
Para las muestras con grupos uniformes, Figuras 2.37 y 2.38, en el caso de la distancia euclídea se tienen diferencias elevadas, ya que cuando hay dos grupos el error máximo es 9 % (Ward), mientras que cuando hay cuatro llega al 64 % (encadenamiento simple). En cuanto a las similaridades, salvo para la de bandas modificadas, las variaciones no son tan sustanciales. Las mejores, de forma global sobre los dos ejemplos, son la similaridad simplicial y la similaridad por bandas.

2.5.2.2. Grupos con distribución asimétrica en al menos una coordenada

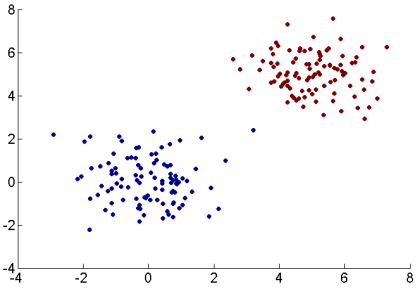
El segundo grupo de muestras está formado por conjuntos de datos formadas por grupos que presentan asimetría en alguna de sus coordenadas. Ésta se introduce por medio de componentes de distribución exponencial. El hecho de introducir asimetría a través de esta función puede distorsionar algunas medidas debido a la aparición de valores



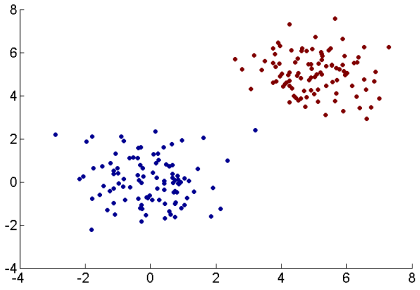
(a) Distancia Euclídea (Ward)



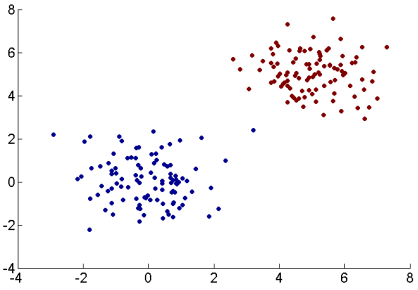
(b) Distancia Euclídea (Simple)



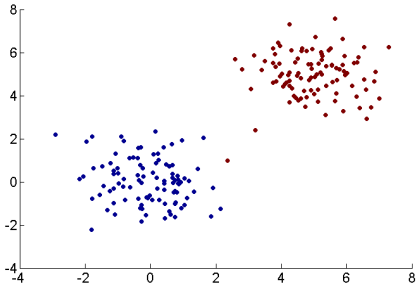
(c) Similaridad de Mahalanobis



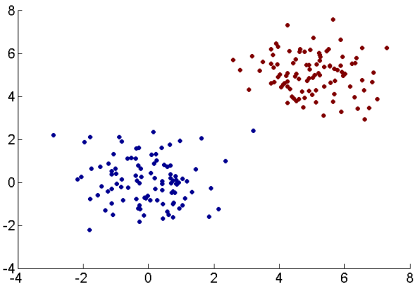
(d) Similaridad por proyecciones



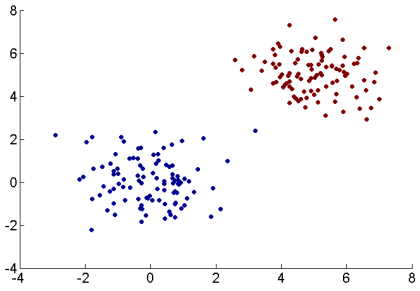
(e) Similaridad de Oja



(f) Similaridad simplicial

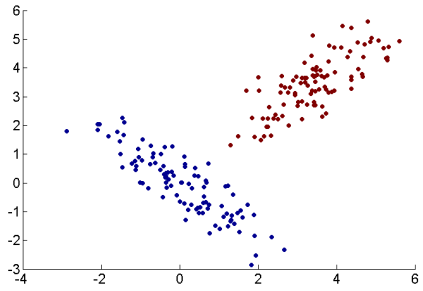


(g) Similaridad por bandas

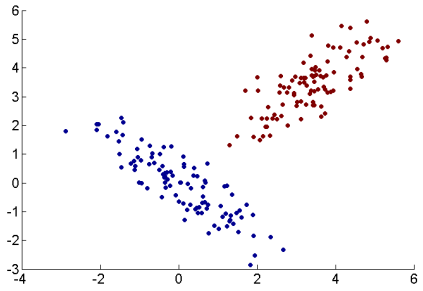


(h) Similaridad por bandas modificada

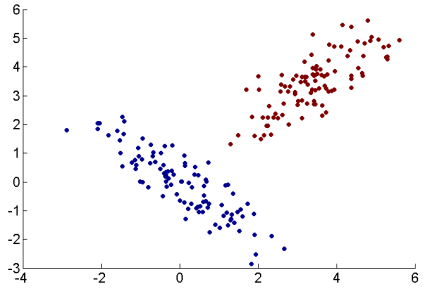
Figura 2.33: *Agrupación para muestra procedente de dos normales de igual matriz de covarianzas y medias distintas.*



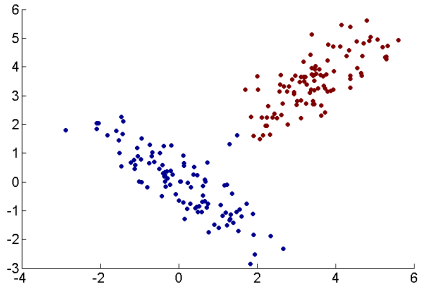
(a) Distancia Euclídea (Ward)



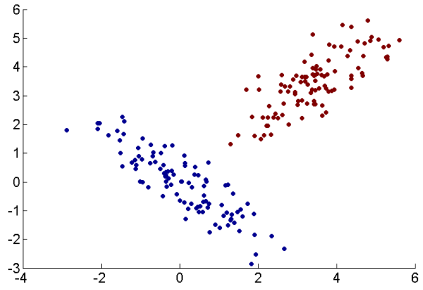
(b) Distancia Euclídea (Simple)



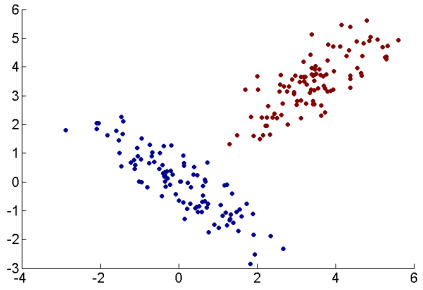
(c) Similaridad de Mahalanobis



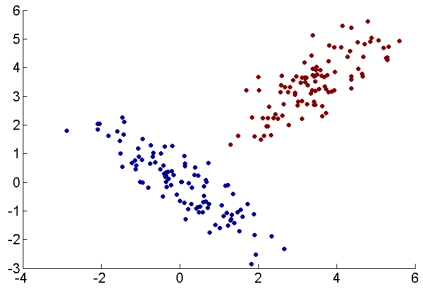
(d) Similaridad por proyecciones



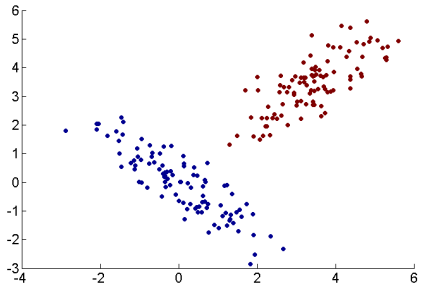
(e) Similaridad de Oja



(f) Similaridad simplicial

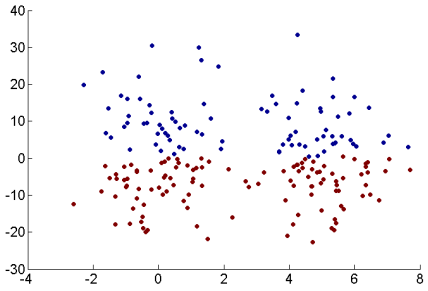


(g) Similaridad por bandas

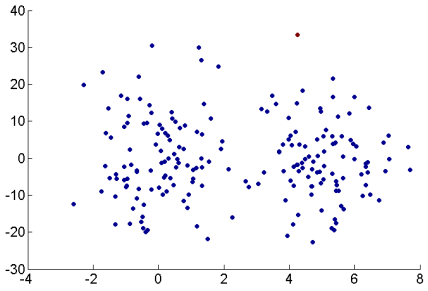


(h) Similaridad por bandas modificada

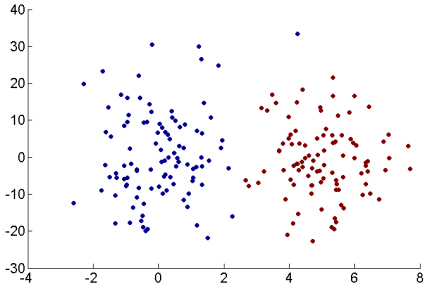
Figura 2.34: *Agrupación para muestra procedente de dos normales con iguales varianzas, correlaciones opuestas y medias distintas.*



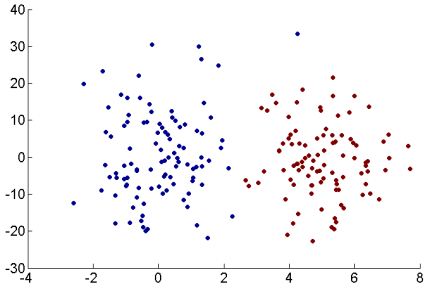
(a) Distancia Euclídea (Ward)



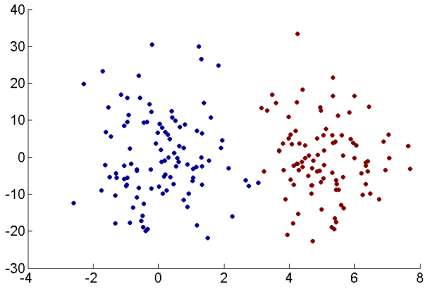
(b) Distancia Euclídea (Simple)



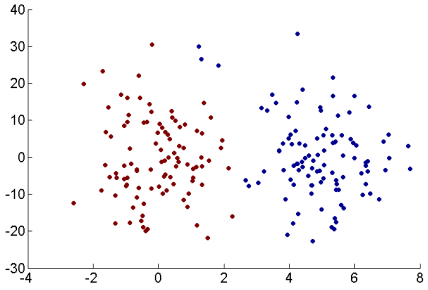
(c) Similaridad de Mahalanobis



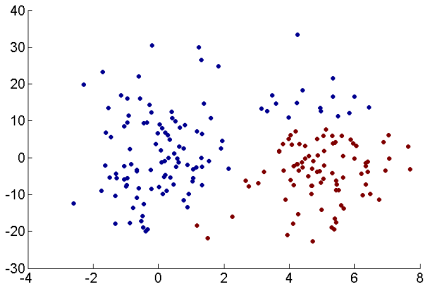
(d) Similaridad por proyecciones



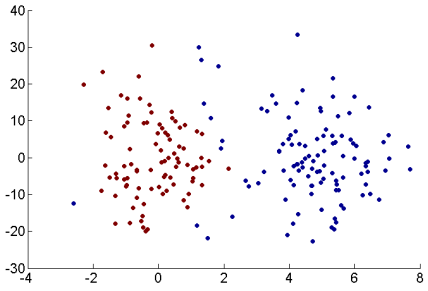
(e) Similaridad de Oja



(f) Similaridad simplicial

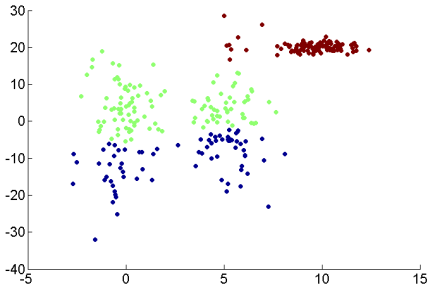


(g) Similaridad por bandas

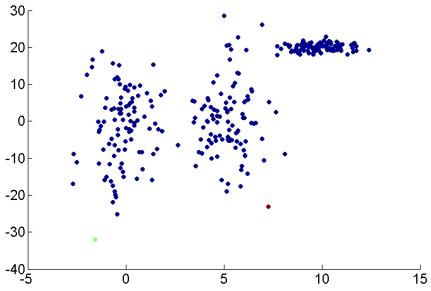


(h) Similaridad por bandas modificada

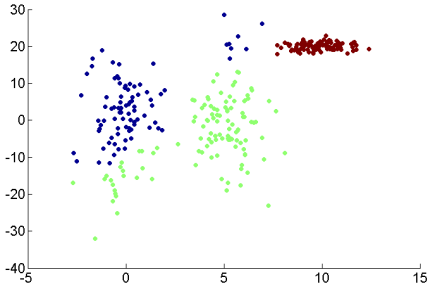
Figura 2.35: *Agrupación para muestra procedente de dos normales con variabilidad elevada en una de sus coordenadas y medias distintas.*



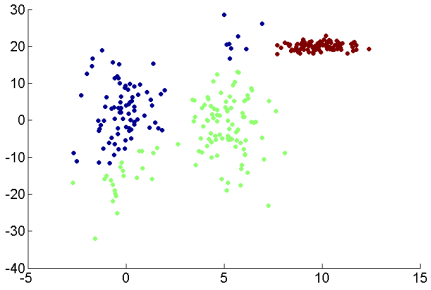
(a) Distancia Euclídea (Ward)



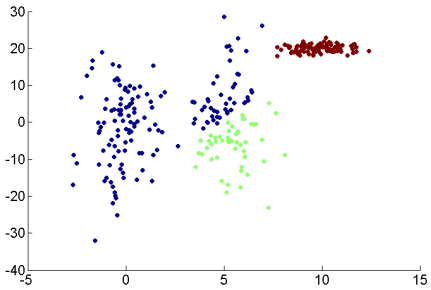
(b) Distancia Euclídea (Simple)



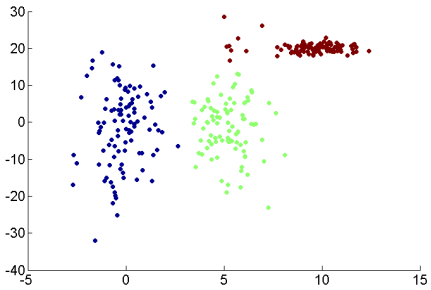
(c) Similaridad de Mahalanobis



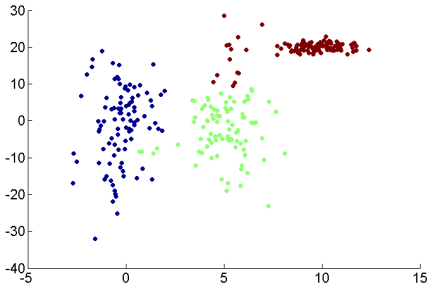
(d) Similaridad por proyecciones



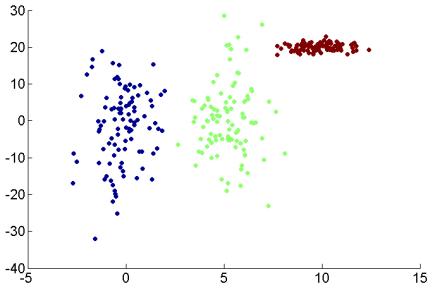
(e) Similaridad de Oja



(f) Similaridad simplicial

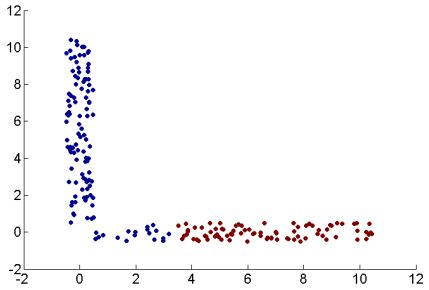


(g) Similaridad por bandas

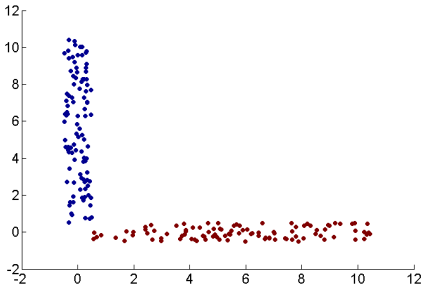


(h) Similaridad por bandas modificada

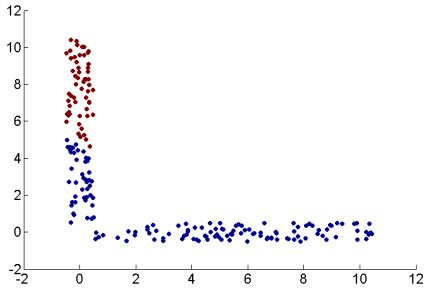
Figura 2.36: *Agrupación para muestra procedente de dos normales con variabilidad elevada en una de sus coordenadas y medias distintas y de otra normal con matriz de covarianza identidad y vector de medias distinto al de los otros grupos.*



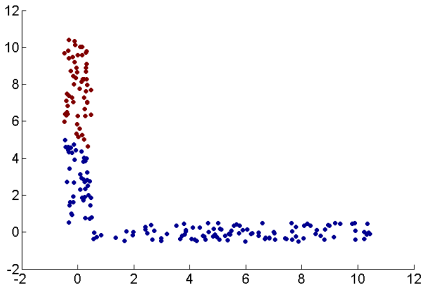
(a) Distancia Euclídea (Ward)



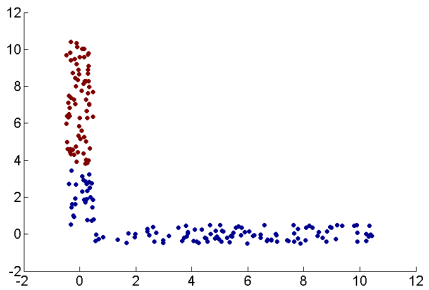
(b) Distancia Euclídea (Simple)



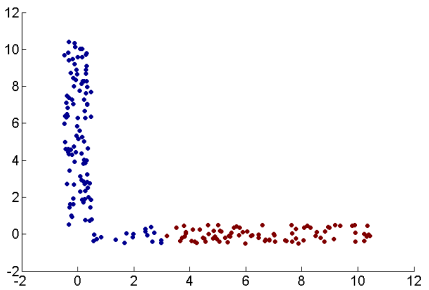
(c) Similaridad de Mahalanobis



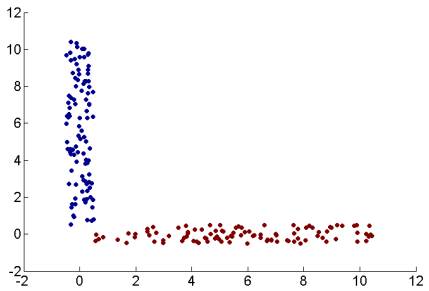
(d) Similaridad por proyecciones



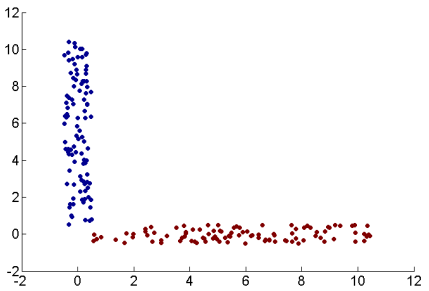
(e) Similaridad de Oja



(f) Similaridad simplicial

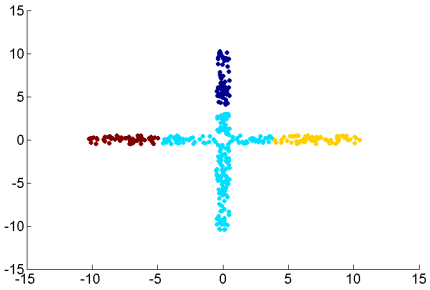


(g) Similaridad por bandas

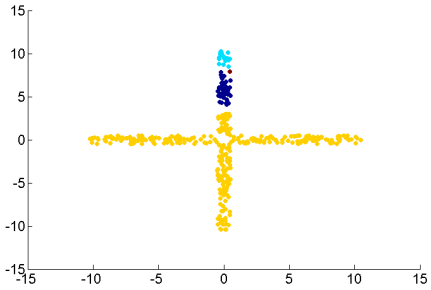


(h) Similaridad por bandas modificada

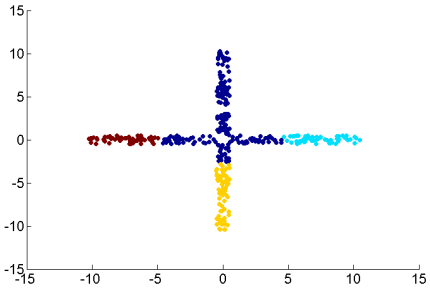
Figura 2.37: *Agrupación para muestra procedente de dos rectángulos uniformes.*



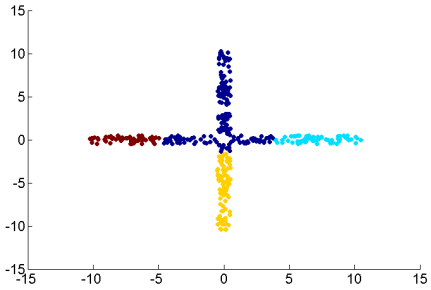
(a) Distancia Euclídea (Ward)



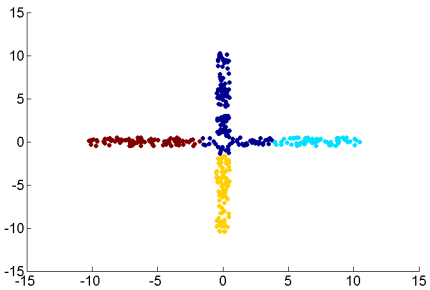
(b) Distancia Euclídea (Simple)



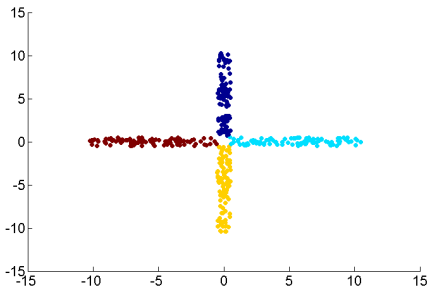
(c) Similaridad de Mahalanobis



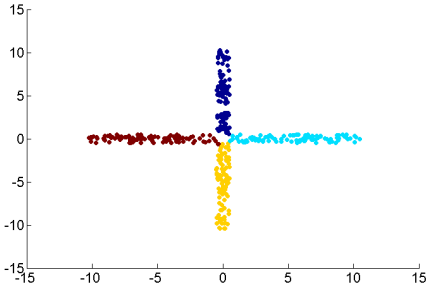
(d) Similaridad por proyecciones



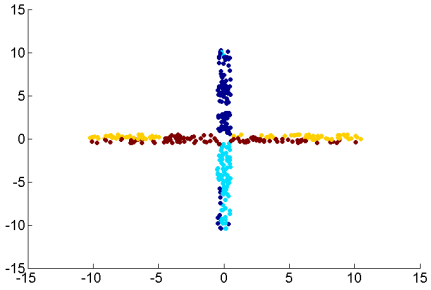
(e) Similaridad de Oja



(f) Similaridad simplicial



(g) Similaridad por bandas



(h) Similaridad por bandas modificada

Figura 2.38: *Agrupación para muestra procedente de cuatro rectángulos uniformes.*

alejados de la mediana de cada variable. Los ejemplos son los siguientes.

Ejemplo 1: Dos grupos de tamaños $n_1 = n_2 = 100$. El primer grupo tiene ambas coordenadas independientes y distribuidas según una exponencial de media igual a uno. El segundo grupo tiene la misma composición pero con signo negativo en ambas variables.

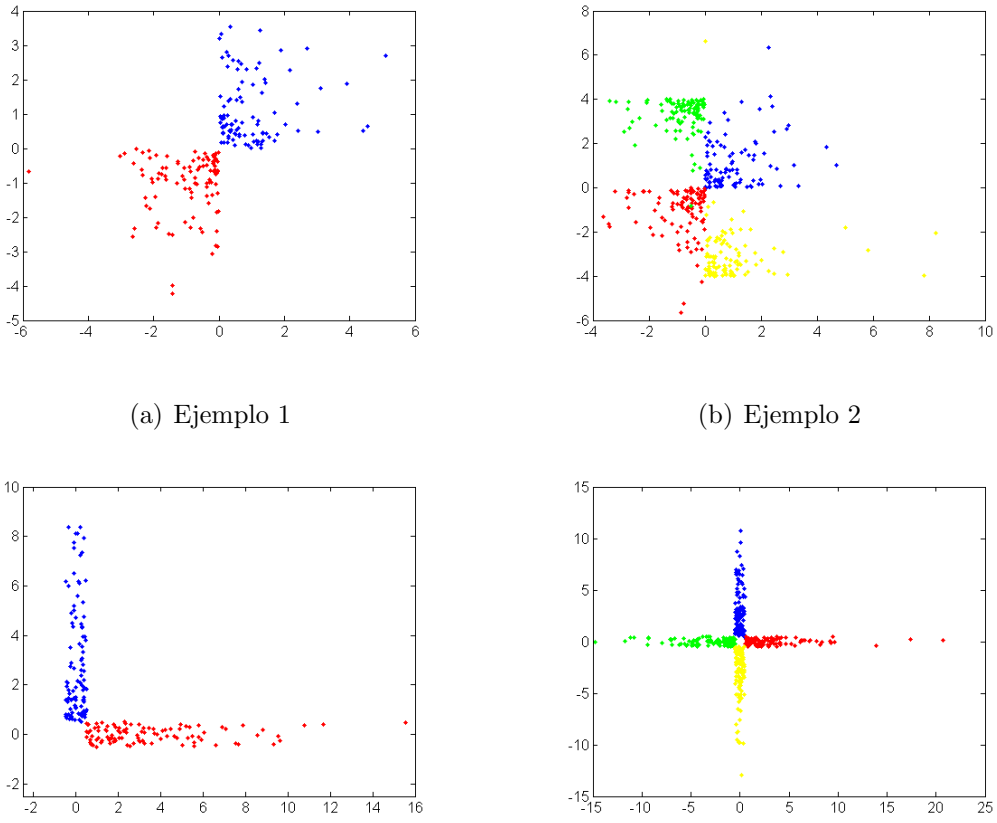
Ejemplo 2: Cuatro grupos de tamaños $n_1 = n_2 = n_3 = n_4 = 100$. Dos de los grupos se simulan conforme a los grupos del Ejemplo 1. El tercero se genera como el primero pero trasladado cuatro unidades hacia abajo (en la segunda coordenada) y el cuarto como el segundo pero trasladado cuatro unidades hacia arriba (segunda coordenada).

Ejemplo 3: Dos grupos de forma aproximadamente rectangular, con una coordenada uniforme y otra exponencial y de tamaños $n_1 = n_2 = 100$. El primer grupo tiene la primera coordenada con distribución U (-0.5, 0.5) y la segunda con exponencial de media igual a 3 y origen en 0.5. El segundo grupo tiene la primera coordenada distribuida según una exponencial de media igual a 3 y origen en 0.5 y la segunda según una U (-0.5, 0.5).

Ejemplo 4: Cuatro grupos de forma aproximadamente rectangular, con una coordenada uniforme y otra exponencial y de tamaños $n_1 = n_2 = n_3 = n_4 = 100$. Dos de estos grupos han sido generados a partir de las distribuciones del ejemplo 3. El tercer grupo tiene la primera coordenada con distribución U (-0.5, 0.5) y la segunda con distribución exponencial de media 3 cambiada de signo y con origen en -0.5 y el cuarto tiene la primera coordenada distribuida según una distribución exponencial de media 3 cambiada de signo y con origen en -0.5 y la segunda según una U (-0.5, 0.5).

La Figura 2.39 contiene los diagramas de puntos de las muestras de los cuatro ejemplos. Puede observarse como el gráfico del Ejemplo 2 (Figura 2.39(d)) no tiene los grupos completamente separables debido a un par de observaciones que aparecen dentro de otros grupos, con lo que el error en este caso no podrá ser en ningún caso igual a cero.

Las Figuras desde 2.40 hasta 2.43 contienen los resultados de agrupamientos para la distancia euclídea y las similaridades. Para la muestra del ejemplo de dos grupos con coordenadas exponenciales, Figura 2.40, todas las funciones tienen errores por debajo del 2 % (casi todas sin errores), salvo la similaridad de Mahalanobis para la que el error alcanza el 34 %. Si son cuatro los grupos, Figura 2.41, los errores se sitúan en torno al 10 %,



(a) Ejemplo 1

(b) Ejemplo 2

(c) Ejemplo 3

(d) Ejemplo 4

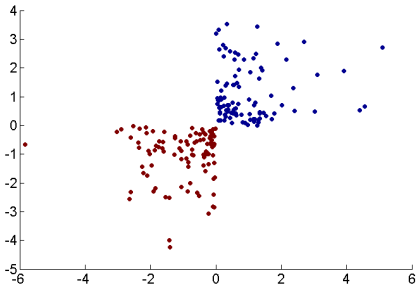
Figura 2.39: *Diagramas de dispersión de los ejemplos con grupos asimétricos.*

exceptuando el encadenamiento simple para la distancia euclídea que se ve afectada por valores extremos (74.8 % de error) y las similaridades de Mahalanobis y por proyecciones con errores en torno al 30 %.

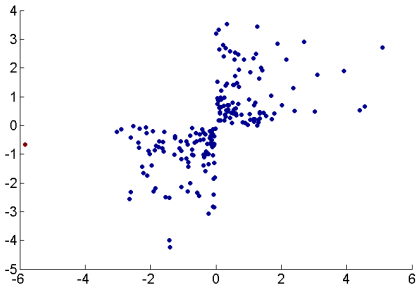
Para las muestras de grupos aproximadamente rectangulares, Figuras 2.42 y 2.43, se tiene que las mejores similaridades son las de bandas y bandas modificada y, en el extremo opuesto, la distancia euclídea, tanto para el método de encadenamiento simple como para el de Ward, aunque éste en menor medida.

2.5.2.3. Grupos con relaciones no lineales entre variables

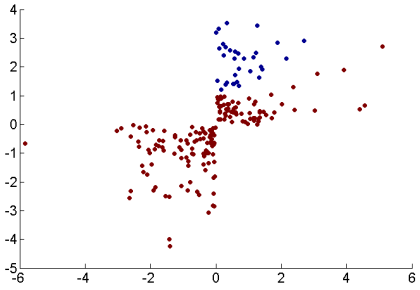
El último grupo de muestras se corresponde a muestras en las que los grupos presentan formas no lineales. Estas formas son circunferencias, anillos y mitades de anillos. La



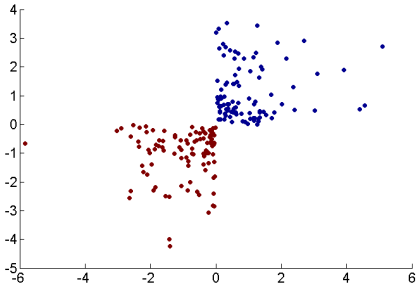
(a) Distancia Euclídea (Ward)



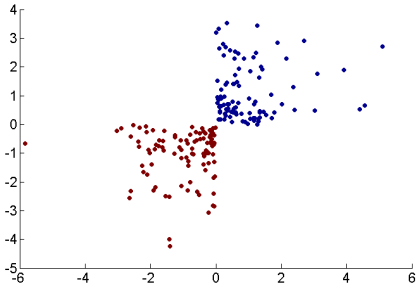
(b) Distancia Euclídea (Simple)



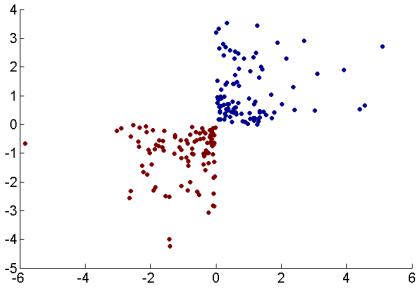
(c) Similaridad de Mahalanobis



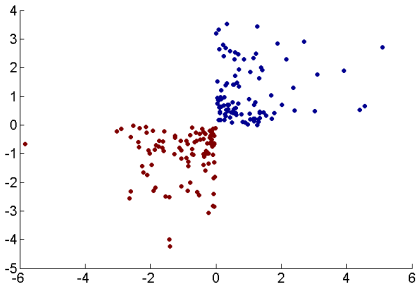
(d) Similaridad por proyecciones



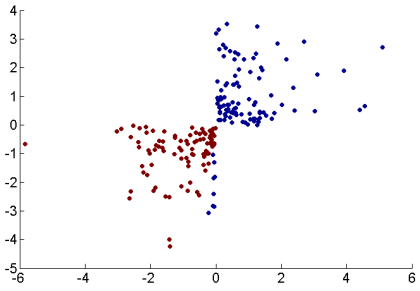
(e) Similaridad de Oja



(f) Similaridad simplicial

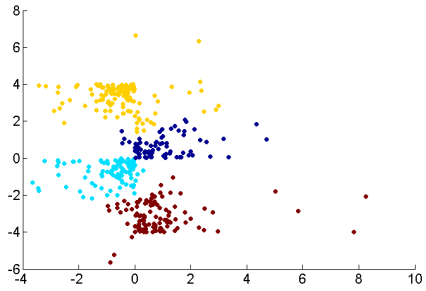


(g) Similaridad por bandas

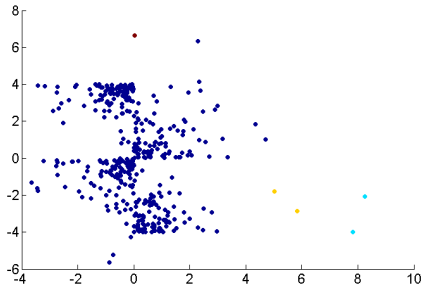


(h) Similaridad por bandas modificada

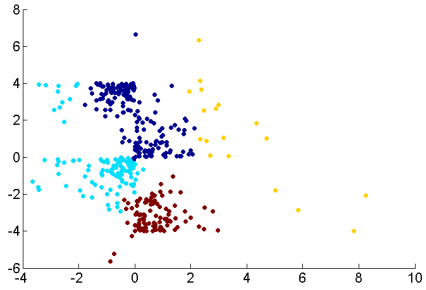
Figura 2.40: *Agrupación para muestra procedente de dos exponenciales, una positiva y otra negativa.*



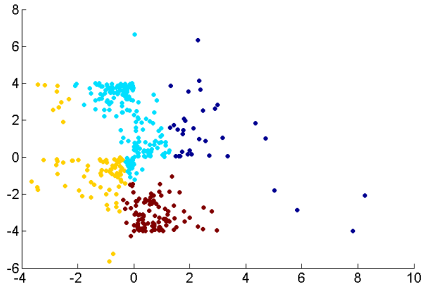
(a) Distancia Euclídea (Ward)



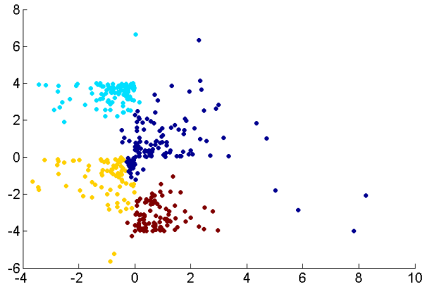
(b) Distancia Euclídea (Simple)



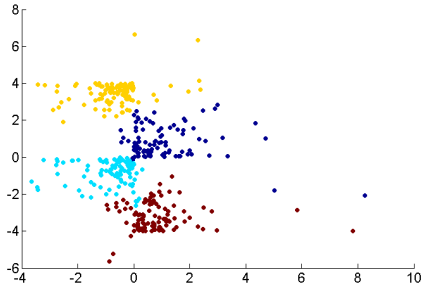
(c) Similaridad de Mahalanobis



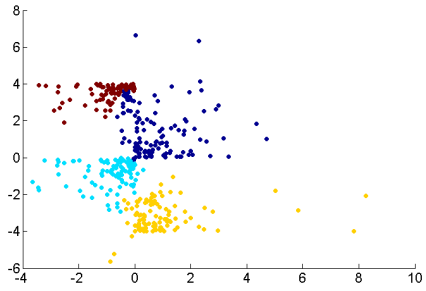
(d) Similaridad por proyecciones



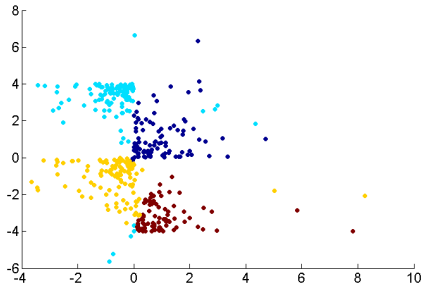
(e) Similaridad de Oja



(f) Similaridad simplicial

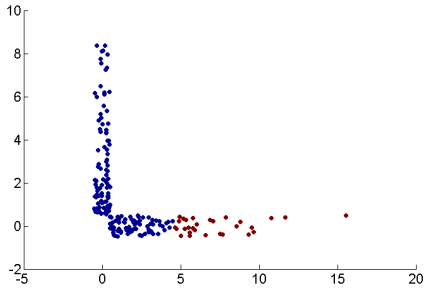


(g) Similaridad por bandas

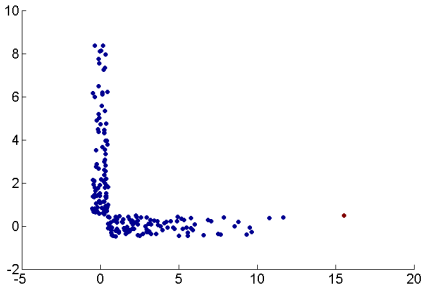


(h) Similaridad por bandas modificada

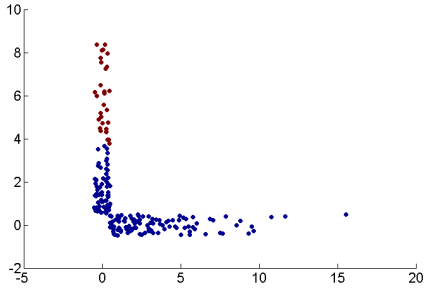
Figura 2.41: *Agrupación para muestra procedente de cuatro exponenciales de signos y orígenes distintos.*



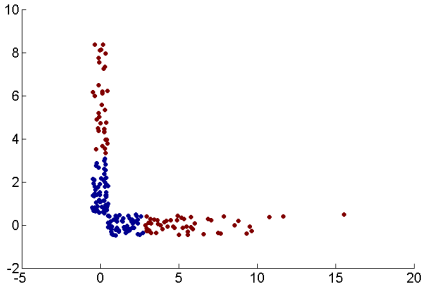
(a) Distancia Euclídea (Ward)



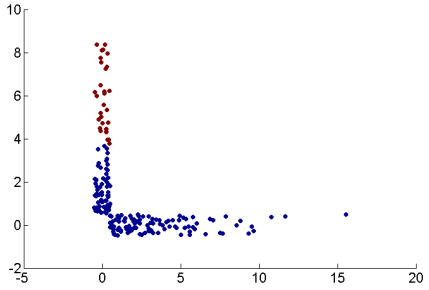
(b) Distancia Euclídea (Simple)



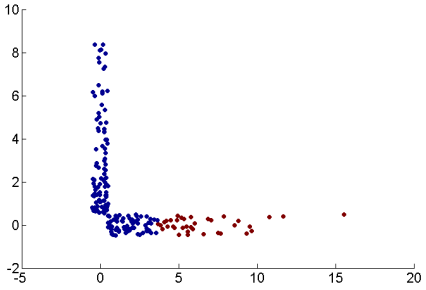
(c) Similaridad de Mahalanobis



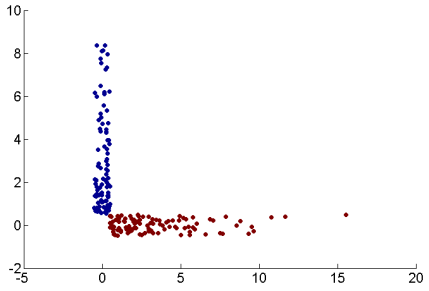
(d) Similaridad por proyecciones



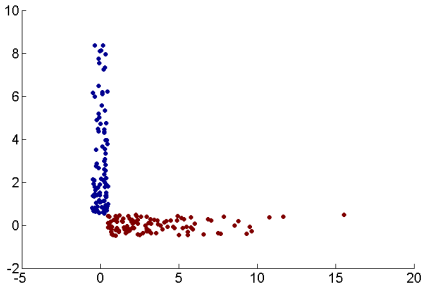
(e) Similaridad de Oja



(f) Similaridad simplicial

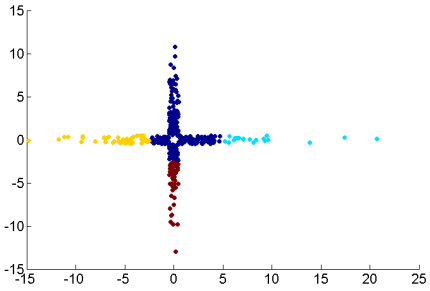


(g) Similaridad por bandas

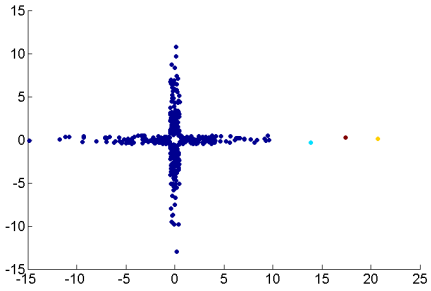


(h) Similaridad por bandas modificada

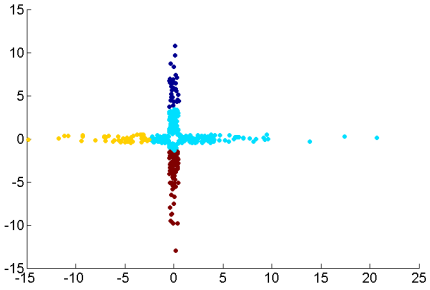
Figura 2.42: *Agrupación para muestra procedente de dos grupos con una coordenada uniforme y otra exponencial.*



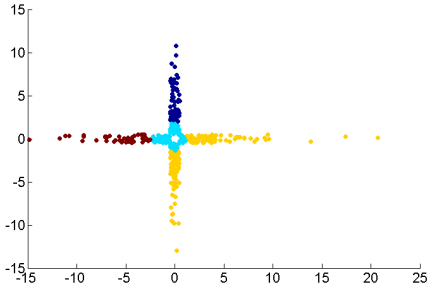
(a) Distancia Euclídea (Ward)



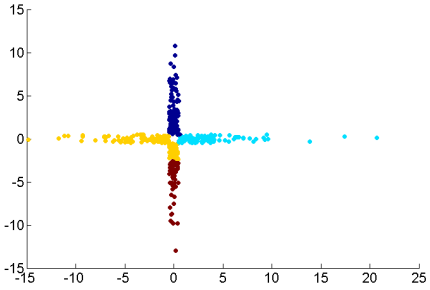
(b) Distancia Euclídea (Simple)



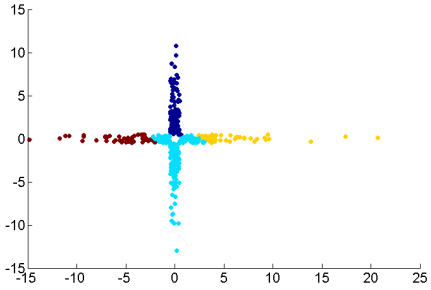
(c) Similaridad de Mahalanobis



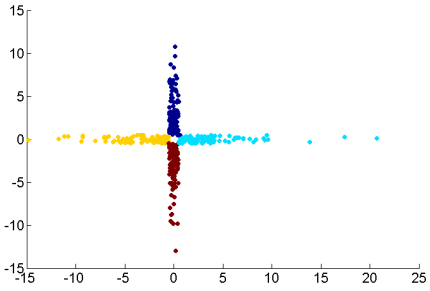
(d) Similaridad por proyecciones



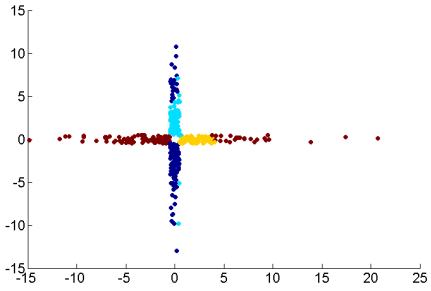
(e) Similaridad de Oja



(f) Similaridad simplicial



(g) Similaridad por bandas



(h) Similaridad por bandas modificada

Figura 2.43: *Agrupación para muestra procedente de cuatro grupos con una coordenada uniforme y otra exponencial.*

generación de las muestras de cada forma se ha llevado a cabo mediante la selección de las observaciones generadas según una distribución uniforme en un cuadrado, que estaban dentro de las regiones de cada forma. A continuación se introducen los cuatro ejemplos.

Ejemplo 1: Dos grupos de tamaños $n_1 = n_2 = 200$. Uno de los grupos tiene distribución uniforme dentro de una circunferencia de radio unidad y el otro distribución uniforme dentro de un anillo circular de radio inferior igual a 1.5 y de radio superior igual a 2.5. Ambas regiones están centradas en el origen.

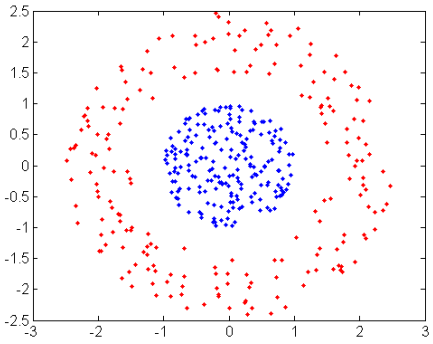
Ejemplo 2: Dos grupos de tamaños $n_1 = n_2 = 200$. Uno de los grupos tiene distribución uniforme dentro de una circunferencia de radio unidad y el otro distribución uniforme dentro de la mitad izquierda de un anillo circular de radio inferior igual a 1.5 y de radio superior igual a 2.5. Ambas regiones están centradas en el origen.

Ejemplo 3: Dos grupos de tamaños $n_1 = n_2 = 100$. Uno de los grupos tiene distribución uniforme dentro de la mitad izquierda de un anillo circular con centro el origen, de radio inferior igual a 1.5 y de radio superior igual a 2.5. El otro grupo es la mitad derecha de otro anillo con los mismos valores de radios y centrado en el punto $(0, 2)'$.

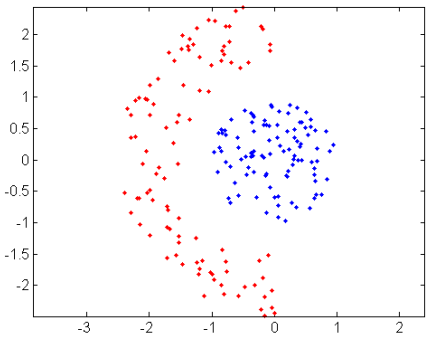
Ejemplo 4: Dos grupos de tamaños $n_1 = n_2 = 100$. Uno de los grupos tiene distribución uniforme dentro de la mitad izquierda de un anillo circular con centro el origen, de radio inferior igual a 1.5 y de radio superior igual a 2.5. El otro grupo es la mitad derecha de otro anillo con los mismos valores de radios y centrado en el punto $(\text{-}0.75, 2)'$.

La Figura 2.44 contiene los diagramas de puntos de los cuatro ejemplos. En tres de los ejemplos es necesario aplicar un clasificador no lineal para separar los grupos completamente. Sólo en el Ejemplo 3 (Figura 2.47) los grupos pueden separarse con una recta.

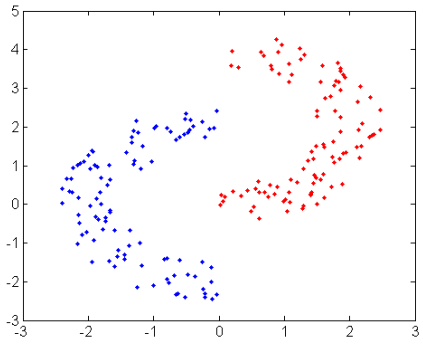
Las agrupaciones resultantes de aplicar el análisis de conglomerados se encuentran representadas en las Figuras 2.45 a 2.48. En el ejemplo de la circunferencia y el anillo alrededor de ésta (Figura 2.45), la única medida que no comete error de agrupamiento es la distancia euclídea con el encadenamiento simple. La mejor similaridad en este ejemplo es la de bandas modificada que obtiene un 27 % de error, valor ligeramente inferior al 30 % obtenido para la distancia euclídea con el método de Ward. Si se toma la mitad



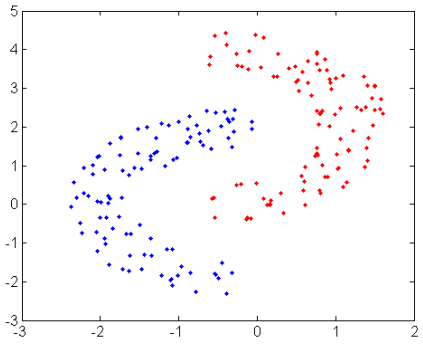
(a) Ejemplo 1



(b) Ejemplo 2



(c) Ejemplo 3

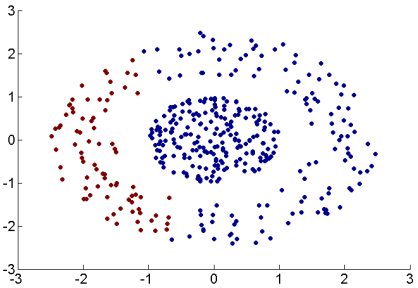


(d) Ejemplo 4

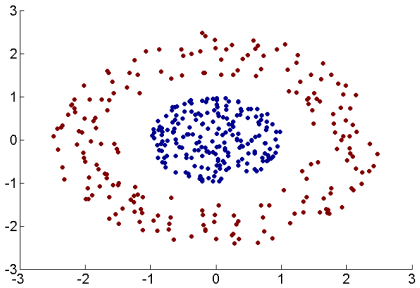
Figura 2.44: *Diagramas de dispersión de los ejemplos con grupos circulares y grupos con formas no lineales.*

del anillo, Figura 2.46, los errores de clasificación disminuyen. El encadenamiento simple para la distancia euclídea agrupa correctamente las observaciones. Las similaridades de Mahalanobis, simplicial y por bandas obtienen errores entre el 13 y el 16 % y la de bandas modificada apenas sufre variación con respecto al ejemplo anterior.

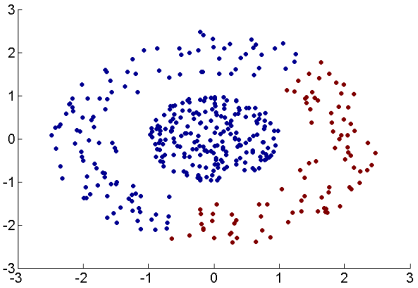
En el Ejemplo 3, Figura 2.47, la distancia euclídea con encadenamiento simple y las similaridades de Mahalanobis y de Oja no cometen errores. La similaridad por proyecciones presenta un 42 % de error, mientras que la de bandas modificada continúa en los niveles de ejemplos anteriores. Por último, en la muestra del Ejemplo 4, Figura 2.48, se tiene que la distancia euclídea con encadenamiento simple sigue sin cometer errores. Las mejores similaridades son la de Mahalanobis, la de bandas y la simplicial.



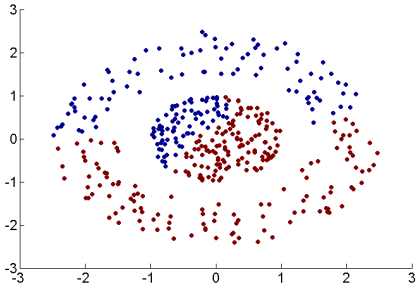
(a) Distancia Euclídea (Ward)



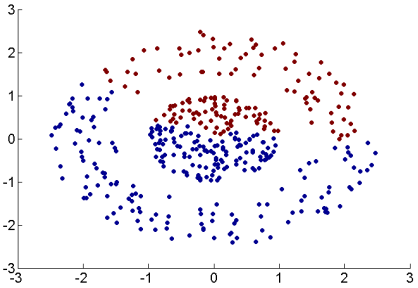
(b) Distancia Euclídea (Simple)



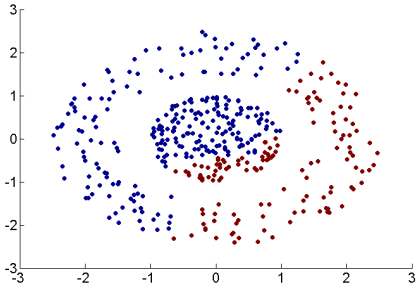
(c) Similaridad de Mahalanobis



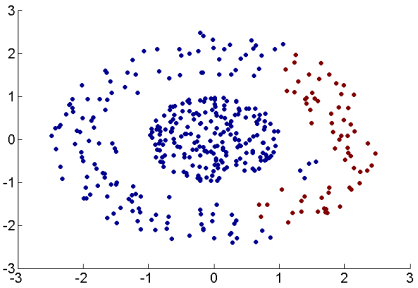
(d) Similaridad por proyecciones



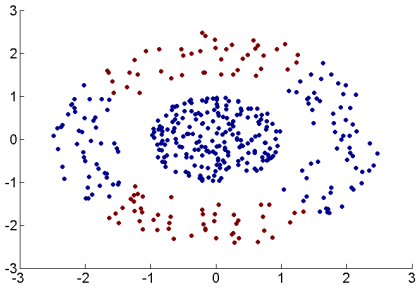
(e) Similaridad de Oja



(f) Similaridad simplicial

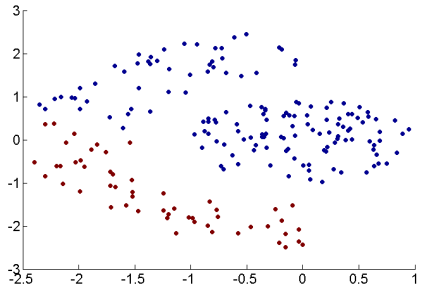


(g) Similaridad por bandas

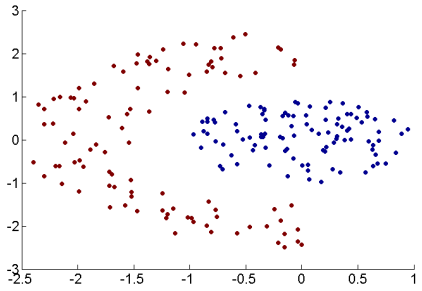


(h) Similaridad por bandas modificada

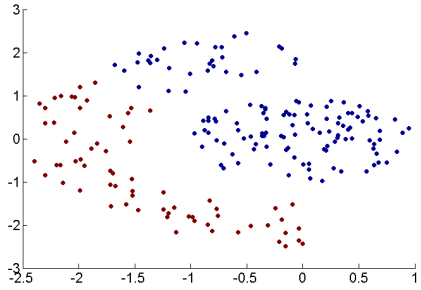
Figura 2.45: *Agrupación para muestra procedente de una uniforme en un círculo y de otra uniforme en un anillo.*



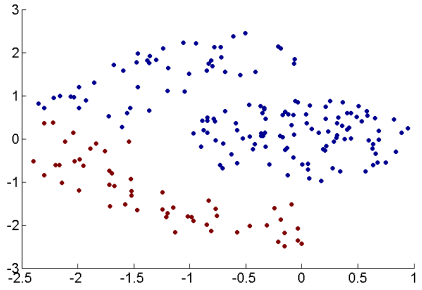
(a) Distancia Euclídea (Ward)



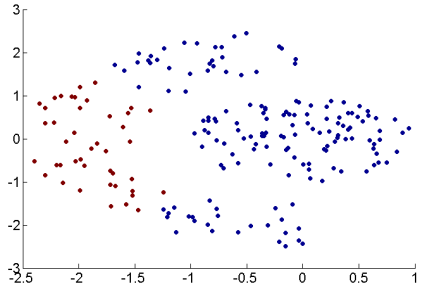
(b) Distancia Euclídea (Simple)



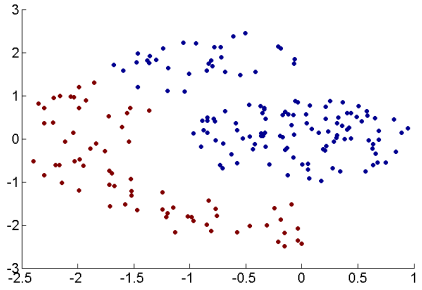
(c) Similaridad de Mahalanobis



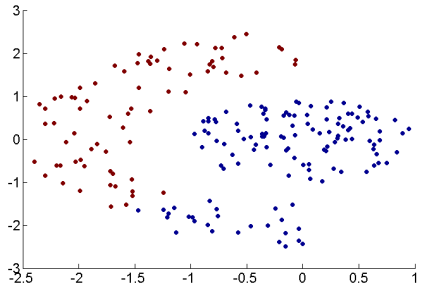
(d) Similaridad por proyecciones



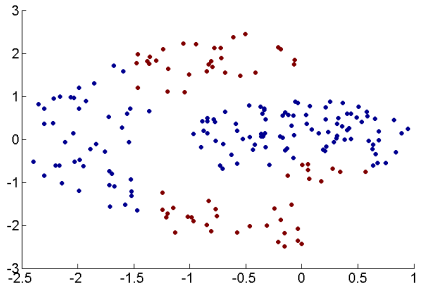
(e) Similaridad de Oja



(f) Similaridad simplicial

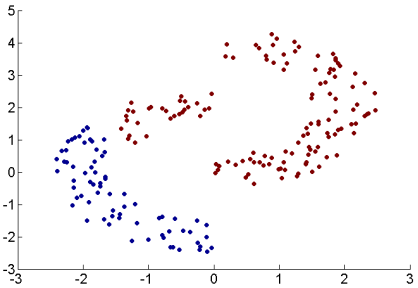


(g) Similaridad por bandas

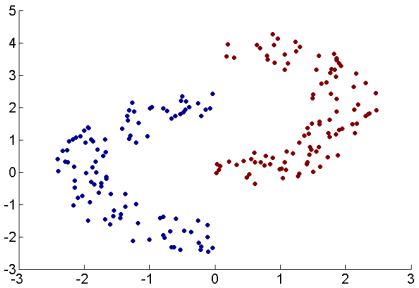


(h) Similaridad por bandas modificada

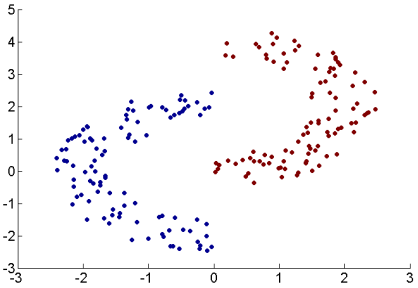
Figura 2.46: *Agrupación para muestra procedente de una uniforme en un círculo y de otra uniforme en una mitad de anillo.*



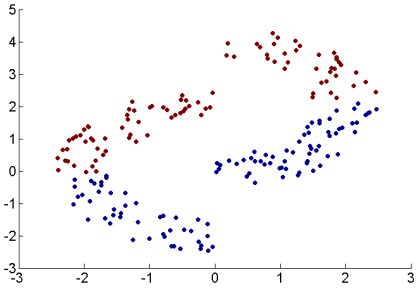
(a) Distancia Euclídea (Ward)



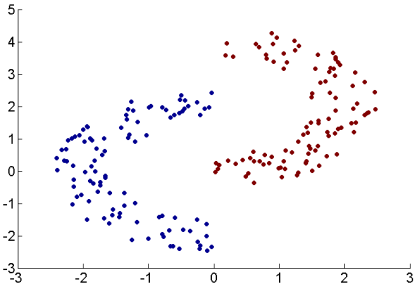
(b) Distancia Euclídea (Simple)



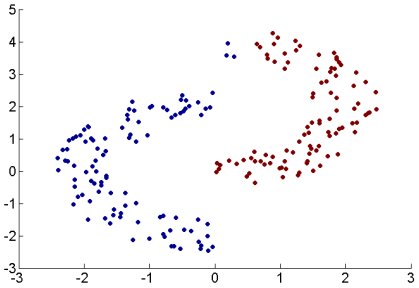
(c) Similaridad de Mahalanobis



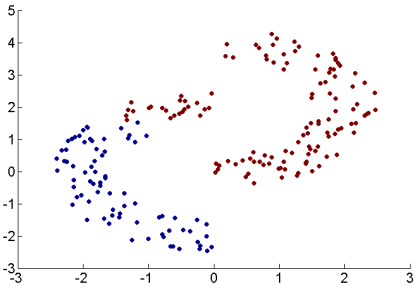
(d) Similaridad por proyecciones



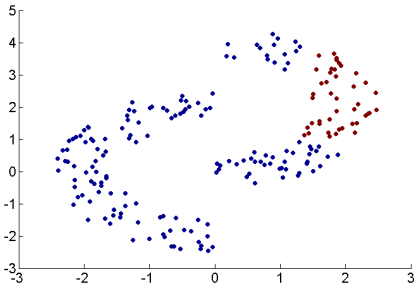
(e) Similaridad de Oja



(f) Similaridad simplicial

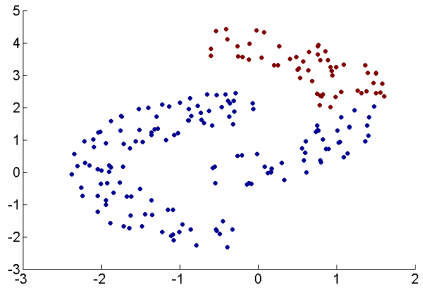


(g) Similaridad por bandas

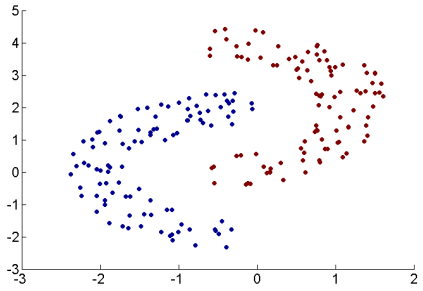


(h) Similaridad por bandas modificada

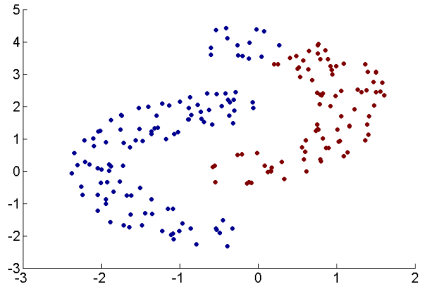
Figura 2.47: *Agrupación para muestra procedente de dos uniformes en mitades de anillos, cuyos centros difieren en la coordenada y.*



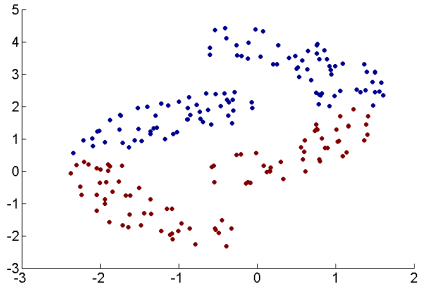
(a) Distancia Euclídea (Ward)



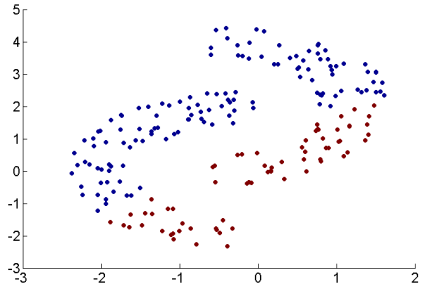
(b) Distancia Euclídea (Simple)



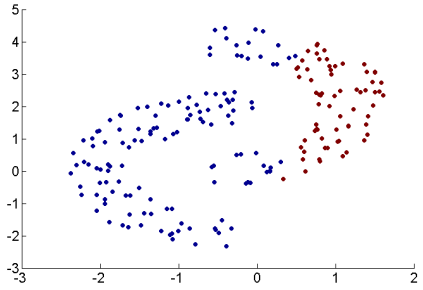
(c) Similaridad de Mahalanobis



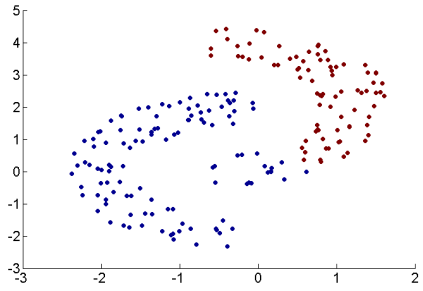
(d) Similaridad por proyecciones



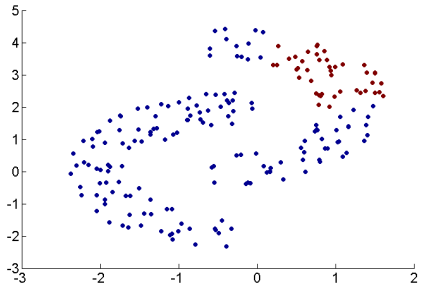
(e) Similaridad de Oja



(f) Similaridad simplicial



(g) Similaridad por bandas



(h) Similaridad por bandas modificada

Figura 2.48: *Agrupación para muestra procedente de dos uniformes en mitades de anillos, cuyos centros difieren en sus dos coordenadas.*

2.5.2.4. Comparación de resultados

En este apartado se recogen y analizan de forma conjunta los errores de clasificación obtenidos por cada función para cada uno de los catorce ejemplos. Esta información se encuentra en la Tabla 2.3.

		Distancia Euclídea		Similaridades					
Tipo	Ejemplo	Ward	Simple	SM	SP	SO	SS	SB	SBM
Simétricos	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	2	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
	3	50.0	49.5	1.5	1.5	1.5	2.5	9.5	5.5
	4	29.7	66.0	10.0	10.0	16.0	3.0	6.0	0.0
	5	9.0	0.0	22.0	22.0	14.5	8.5	0.0	0.0
	6	26.5	64.0	24.8	19.8	12.8	0.2	0.5	26.8
Asimétricos	1	0.0	0.5	34.0	0.0	0.0	0.0	0.0	4.0
	2	11.5	74.8	29.5	29.5	9.0	9.8	10.3	6.8
	3	35.0	49.5	35.0	42.0	35.0	31.5	0.0	0.0
	4	45.8	74.8	37.3	45.8	13.5	25.3	0.7	13.3
No lineales	1	30.3	0.0	28.3	48.0	49.5	40.5	33.8	27.0
	2	23.5	0.0	16.0	23.5	28.5	16.0	13.0	26.5
	3	14.5	0.0	0.0	42.0	0.0	1.5	11.5	28.5
	4	22.0	0.0	7.0	48.0	40.0	16.5	8.5	28.5
	Media	21.3	27.1	17.6	23.8	15.8	11.1	6.7	11.9

Tabla 2.3: *Errores de clasificación del análisis de conglomerados.*

Atendiendo a la media global sobre todos los ejemplos la función que mejor resultados de agrupamiento obtiene es la similaridad por bandas, seguida de la simplicial y la de bandas modificada. El error máximo obtenido para la similaridad por bandas es del 33.8 %. Sólo la similaridad por bandas modificada tiene un error máximo menor que éste.

La distancia euclídea, tanto para el encadenamiento simple como para el de Ward, pre-

senta valores medios que mejoran únicamente la media de la similaridad por proyecciones. Además presenta cierta variabilidad frente al método elegido, es decir, para algunos ejemplos el encadenamiento simple es mucho peor que el de Ward y viceversa. Esto muestra la dependencia entre el error de agrupamiento, la elección del método de cálculo de las distancias entre grupos y la forma de las agrupaciones existentes. Lo que puede respresentar un problema si no se tiene información a priori de la forma de éstas.

Por último se tiene que el peor funcionamiento de las similaridades se obtiene cuando existen grupos de forma no lineal. En ese caso, la distancia euclídea se muestra superior. En especial, si se emplea el método de Ward. Sobre muestras formadas por grupos simétricos, en general, los errores para las similaridades son menores. Este comportamiento se acentúa cuando una de las variables presenta una variabilidad sustancialmente mayor que la otra (Ejemplos 3 a 6 de la categoría de simétricos). Para grupos con asimetría en alguna de sus variables la distancia euclídea con el método de Ward obtienen tasas de error elevadas debido a que forma grupos con sólo una observación (atípica). Para este tipo de ejemplos las dos mejores funciones son la similaridad por bandas y por bandas modificada con tasas de error de 2.75 % y 6.03 % respectivamente.

Capítulo 3

Distancias basadas en similaridades

Resumen

En este capítulo se introduce un nuevo tipo de distancias, basadas en la idea de profundidad, que se obtienen como extensión de las similaridades definidas en el Capítulo 2. El objetivo del capítulo es proponer y aplicar nuevas métricas que sean de utilidad para problemas multivariantes como los métodos de clasificación. Como sucede con las similaridades del capítulo anterior, las funciones que se introducen tienen la capacidad de adaptación no sólo a la forma de la distribución generadora de las observaciones, sino también a la posición que éstas ocupan en el espacio, de ahí que puedan ser consideradas distancias en sentido estadístico. El paso de las similaridades a las distancias se realiza a través de transformaciones cuya mayor dificultad radica en el cumplimiento de la propiedad triangular. Aquí se proponen algunas transformaciones para las similaridades estudiadas previamente y se comprueba que cumplen las propiedades de distancia. La transformación a aplicar no es genérica, depende de la forma funcional de la similaridad que se quiera adaptar. Para ilustrar el comportamiento de las distancias y para mostrar su capacidad de adaptación para distribuciones de formas diferentes, se presentan, como en el capítulo anterior, gráficos con curvas de nivel que muestran la distancia entre cualquier punto del plano y otro punto fijado previamente. Por último, para justificar la aplicación y la utilidad de estas distancias, se presentan los resultados de su aplicación al análisis de conglomerados. Estos resultados se obtienen mediante el algoritmo de k-medias, sobre

el que se propone una modificación que mejora el rendimiento del algoritmo para las distancias basadas en profundidad. Con éste se tiene, que las similaridades simplical, por bandas y por bandas modificadas mejoran los resultados obtenidos por la distancia euclídea, tanto para el algoritmo modificado, como para el algoritmo de k-medias.

3.1. Distancias basadas en similaridades

En esta sección se proponen las transformaciones con que se consigue el paso de similaridades a distancias. Aunque algunas de las similaridades propuestas en el capítulo anterior están definidas a partir de funciones que ya son distancias (por ejemplo la de Mahalanobis), se continúa con ellas, pues uno de los objetivos principales es ilustrar que con las funciones de profundidad pueden construirse tanto similaridades como distancias.

Las distancias basadas en profundidad que se proponen a continuación se definen a partir de la clasificación de funciones de profundidad hecha en Zuo y Serfling (2000a). Tal y como se hizo en el capítulo anterior, se puede diferenciar entre las similaridades procedentes de profundidades denotadas como tipo A y las denotadas como tipos B y C.

Se tratarán todas las similaridades propuestas en el capítulo anterior: de Mahalanobis, por proyecciones y de Oja, procedentes de profundidades de tipo B y C, y la simplicial, por bandas y por bandas modificada procedentes de profundidades de tipo A. Como se comentó en el capítulo anterior, para las tres primeras el rango de posibles valores que la similaridad puede tomar no depende de los puntos x e y, es decir, fijando uno de los dos puntos el rango de la similaridad coincide con el intervalo $[0, 1)$. Sin embargo, para las otras tres similaridades, fijado uno de los puntos el rango depende de ese punto, es decir, el rango para dos puntos cualesquiera x e y es el intervalo $[0, \text{mín}\,(P(x), P(y)))$, donde $P(x)$ es la profundidad del punto x calculada a partir de la función de profundidad de la que procede la similaridad. Debido a estas diferencias las distancias basadas en profundidad se definen dependiendo de la forma funcional y del rango de posibles valores de las similaridades.

Se comienza proponiendo la transformación para las similaridades basadas en distancias y en medidas de atipicidad (tipos B y C). Es bien conocido que, para cualquier distancia d, la función $d/(1+d)$ es también una distancia. Basado en este resultado, y debido a la expresión de este tipo de similaridad, $S(x,y) = 1/(1+A(x,y))$, se tiene que $1 - S(x,y)$ es una distancia si la medida de atipicidad $A(x,y)$ también lo es. Por definición estas tres distancias son continuas (incluso si la distribución no lo es), por lo que verifican:

(i) $D(x, y; F) = 0$ sí y sólo sí $x = y$

(ii) $D(x, y; F) \geq 0 \ \forall x \in \mathbb{R}^d$

(iii) $D(x, y; F) = D(y, x; F) \ \forall x, y \in \mathbb{R}^d$.

Por lo tanto, bastaría con probar la propiedad triangular para las tres funciones.

Observación 3.1 *La similaridad de Mahalanobis se ha definido empleando el cuadrado de la distancia. Si en vez de usar el cuadrado se usa la distancia, se tiene que* $DM(x, y) = 1 - SM(x, y)$ *es también una distancia. Las ordenaciones de los puntos según la distancia sobre un punto fijo no varían al tomar la distancia en vez del cuadrado.*

La similaridad de Oja, bajo la transformación $1 - SO(x, y)$ no es una distancia, ya que la desigualdad triangular no se cumple para todas las funciones de distribución. Esto puede verse mediante un contraejemplo: suponiendo que la función de distribución sobre la que se calcula la similaridad es degenerada en un punto $t \in \mathbb{R}^d$ y que se tienen los puntos $x, y, z \in \mathbb{R}^d$ dispuestos como en la Figura 3.1, entonces $1 - SO(x, z) > 1 - SO(x, y) + 1 - SO(y, z)$, ya que $SO(x, y) = 1$ (están alineados) y el volumen del símplice $S[x, z, t]$ es mayor que $S[y, z, t]$.

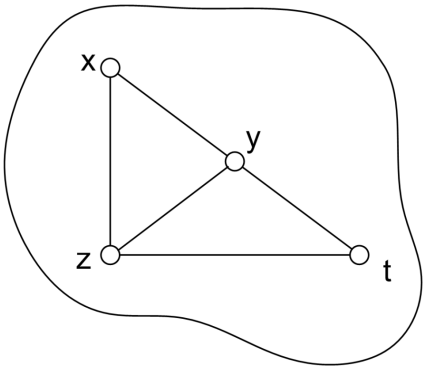


Figura 3.1: *Contraejemplo similaridad de Oja.*

Proposición 3.1 *La similaridad por proyecciones verifica que* $1 - SP(x, y)$ *es una distancia.*

Demostración. Para demostrar que es distancia basta con ver que verifica la propiedad triangular. La medida de atipicidad de la similaridad entre dos puntos x y z se ha definido como

$$A(x,z;F) = \sup_{\|u\|=1} \frac{|u'x - u'z|}{MEDA(u'X)}.$$

Sea $v \in \mathbb{R}^d$ el vector con norma igual a 1, tal que

$$\sup_{\|u\|=1} \frac{|u'x - u'z|}{MEDA(u'X)} = \frac{|v'x - v'z|}{MEDA(v'X)},$$

entonces sumando y restando la cantidad $v'y$ dentro del valor absoluto del numerador se tiene que

$$\frac{|v'x - v'z|}{MEDA(v'X)} = \frac{|v'x - v'y + v'y - v'z|}{MEDA(v'X)} \leq \frac{|v'x - v'y|}{MEDA(v'X)} + \frac{|v'y - v'z|}{MEDA(v'X)}$$

y, debido a que

$$\frac{|v'x - v'y|}{MEDA(v'X)} \leq \sup_{\|u\|=1} \frac{|u'x - u'y|}{MEDA(u'X)} = A(x,y;F)$$

y

$$\frac{|v'y - v'z|}{MEDA(v'X)} \leq \sup_{\|u\|=1} \frac{|u'y - u'z|}{MEDA(u'X)} = A(y,z;F),$$

se obtiene que $A(x,z;F) \leq A(x,y;F) + A(y,z;F)$, por lo que A es distancia y por tanto, $DP(\cdot,\cdot;F)$ también lo es. ■

Tras el análisis de este tipo de distancias basadas en profundidad, se pasa a estudiar las similaridades de tipo A: las basadas en probabilidades de pertenencia a conjuntos aleatorios. Como se ha comentado anteriormente el valor de estas medidas fijado uno de los dos puntos depende del valor de la profundidad de dicho punto. Debido a esto hay que tener en cuenta la proximidad de los puntos al centro introduciendo esta información en la expresión que define las distancias.

Para ilustrar la idea de la transformación que se introduce a continuación, se toma la profundidad y la similaridad simplicial. En el caso bidimensional, la similaridad busca el conjunto de triángulos que contienen a los puntos x e y. Así pues, denotando por A_x y A_y, respectivamente, a los conjuntos de todos esos triángulos para x y para y, se obtiene que la similaridad simplicial es igual a la probabilidad de la intersección de ambos conjuntos.

Pero, como se desea tener una medida amplia que no dependa de las profundidades de cada punto, es necesario cuantificar un límite superior a esta intersección. Si x se encuentra lejos de y, entonces pocos triángulos que estén en A_x estarán en A_y, y viceversa. Mientras que si los puntos x e y se encuentran próximos, habrá entonces muchos triángulos de A_x que estén en A_y, y recíprocamente. En el límite, cuando ambos puntos sean iguales, la intersección de los dos conjuntos de triángulos será igual a dichos conjuntos y por tanto a su unión. De ahí que, para obtener una medida independiente de la posición de los puntos pero dependiente de la distancia que los separa se propone una transformación que tenga en cuenta ambas cantidades: la intersección y la unión.

Por lo tanto, dada una similaridad por profundidad $S(x, y; F)$ basada en probabilidades de conjuntos aleatorios, que es extensión de una función de profundidad $P(x; F)$ (funciones de profundidad de tipo A y sus similaridades), la función que se propone como distancia entre puntos es $1 - SP(x, y, F) / (P(x, F) + P(y, F) - SP(x, y, F))$, es decir, $1 - Pr(A_x \cap A_y) / Pr(A_x \cup A_y)$, donde A_x y A_y son los conjuntos aleatorios de las formas geométricas en que se basa la profundidad, que contienen respectivamente a los puntos x e y. El siguiente resultado recoge el cumplimiento de la propiedad triangular para las distancias así definidas.

Observación 3.2 *Sean $(\Omega, \mathbf{A}, Pr)$ un espacio de probabilidad y A, B y C tres elementos cualesquiera de $\mathbf{A}$. Se tiene que*

$$\frac{Pr(A \Delta C)}{Pr(A \cup C)} \leq \frac{Pr(A \Delta B)}{Pr(A \cup B)} + \frac{Pr(B \Delta C)}{Pr(B \cup C)},$$

donde $A \Delta C$ es $(A \setminus C) \cup (C \setminus A)$.

Demostración. Se comienza con la parte izquierda de la desigualdad. Se introduce la probabilidad de un conjunto tanto en el numerador como en el denominador. En este caso, el cociente es menor o igual que 1 y el sumando que se introduce es mayor o igual que 0. Es sencillo comprobar que el resultado es mayor que el cociente anterior. El conjunto cuya probabilidad se introduce en este paso es $B \setminus (A \cup C)$. Entonces se tiene que

$$\frac{Pr(A \Delta C)}{Pr(A \cup C)} \leq \frac{Pr(A \Delta C) + Pr(B \setminus (A \cup C))}{Pr(A \cup C) + Pr(B \setminus (A \cup C))}.$$

Por un lado, en el denominador se tiene que $(A \cup C) \sqcup (B \setminus (A \cup C)) = A \cup B \cup C$, donde $\sqcup$ es la unión disjunta. Por lo que el denominador puede reemplazarse por la expresión $P(A \cup B \cup C)$. Por otro lado, es sencillo probar que

$$(A \Delta C) \sqcup (B \setminus (A \cup C)) \subset (A \Delta B) \cup (B \Delta C),$$

por lo que

$$\frac{Pr(A \Delta C)}{Pr(A \cup C)} \leq \frac{Pr((A \Delta C) \sqcup (B \setminus (A \cup C)))}{Pr(A \cup B \cup C)} \leq \frac{Pr(A \Delta B)}{Pr(A \cup B \cup C)} + \frac{Pr(B \Delta C)}{Pr(A \cup B \cup C)}$$

y, debido a que $Pr(A \cup B \cup C)$ es mayor o igual que $P(A \cup B)$ y que $Pr(B \cup C)$, se obtiene que

$$\frac{Pr(A \Delta C)}{Pr(A \cup C)} \leq \frac{Pr(A \Delta B)}{Pr(A \cup B)} + \frac{Pr(B \Delta C)}{Pr(B \cup C)}. \blacksquare$$

Por último, las tres similaridades por profundidad de este tipo verifican por definición la propiedad de simetría. Según el teorema 2.3 (Similaridad Simplicial) y para las funciones de distribución que verifican sus supuestos, se tiene que la distancia definida a continuación es igual a cero si, y sólo si, los dos puntos son iguales. Por lo que la siguiente proposición queda demostrada.

Proposición 3.2 *Sea $D \subseteq \mathbb{R}^d$ un conjunto abierto y sea $F : D \to \mathbb{R}^+$ una función de distribución continua. Entonces la función $DS(x, y; F) = 1 - SS(x, y; F) / (PS(x; F) + PS(y; F) - SS(x, y; F))$ (o distancia simplicial) es una distancia en $\mathbb{R}^d$.*

Para las distancias por bandas y por bandas modificada no son válidos de forma directa los resultados 2.8 y 2.13, debido a que en el paso a distancia no se estandariza la similaridad en sí, sino que se estandariza cada uno de los sumandos. Para la similaridad por bandas, con el mismo razonamiento y las mismas hipótesis del Teorema 2.8, se comprueba de forma inmediata que cada sumando $SB^{b)}$ verifica que $SB^{b)}(x, y; F) = SB^{b)}(x, x; F)$ si y sólo si $x = y$, lo que implica que la distancia entre x e y es igual a cero si y sólo si $x = y$. La Observación 3.2 asegura que la transformación sobre cada sumando verifica la propiedad triangular. Con estos resultados queda demostrada la siguiente proposición.

Proposición 3.3 *Sea $D \subseteq \mathbb{R}^d$ un conjunto abierto y sea $F : D \to \mathbb{R}^+$ una función de distribución continua. Entonces la función $DB(x, y; F, B) = \sum_{b=2}^{B} \left(1 - SB^{b)}(x, y; F) / \left(PB^{b)}(x; F) + PB^{b)}(y; F) - SB^{b)}(x, y; F)\right)\right)$, donde $PB^{b)}(x; F) = Pr(x \in R(X_1, \ldots, X_b))$ y $SB^{b)}(x, y; F) = Pr(x, y \in R(X_1, \ldots, X_b))$, es una distancia en $\mathbb{R}^d$.*

Por último, se define la distancia por bandas modificada.

Definición 3.1 *Sea F una función de distribución continua en $\mathbb{R}^d$. Dados dos puntos x e y en $\mathbb{R}^d$. Se define el componente k-ésimo para bandas formadas por b puntos de la profundidad por bandas modificada como:*

$$PBM^{b,k)}\left(x^{k)}; F\right) = Pr\left(\text{mín}\left(X_1^{k)}, X_2^{k)}, \ldots, X_b^{k)}\right) \leq x^{k)} \leq \text{máx}\left(X_1^{k)}, X_2^{k)}, \ldots, X_b^{k)}\right)\right)$$

y el componente k-ésimo para bandas formadas por b puntos de la similaridad por bandas modificada como:

$$SBM^{b,k)}\left(x^{k)}, y^{k)}; F\right) = Pr\left(\text{mín}\left(X_1^{k)}, X_2^{k)}, \ldots, X_b^{k)}\right) \leq x^{k)}, y^{k)} \leq \text{máx}\left(X_1^{k)}, X_2^{k)}, \ldots, X_b^{k)}\right)\right).$$

Proposición 3.4 *Sea $D \subseteq \mathbb{R}^d$ un conjunto abierto y sea $F : D \to \mathbb{R}^+$ una función de distribución continua. Entonces la función*

$$DBM(x, y; F, B) = \frac{1}{d} \sum_{b=2}^{B} \sum_{k=1}^{d} \left(1 - \frac{SBM^{b,k)}\left(x^{k)}, y^{k)}; F\right)}{PBM^{b,k)}\left(x^{k)}; F\right) + PBM^{b,k)}\left(y^{k)}; F\right) - SBM^{b,k)}\left(x^{k)}, y^{k)}; F\right)}\right)$$

(o distancia por bandas modificada) es una distancia en $\mathbb{R}^d$.

Demostración. La demostración está también basada en la Observación 3.2. Gracias a este resultado se tiene que cada sumando verifica la desigualdad triangular, por lo que el sumatorio doble también la satisface. Para esta distancia hay que demostrar que efectivamente la distancia entre dos puntos es igual a cero si, y sólo si, los dos puntos son iguales. Si $x = y$ por definición se verifica que la distancia es nula. La implicación opuesta es cierta ya que para cualquier b, si la suma

$$\sum_{k=1}^{d} \left(1 - \frac{SBM^{b,k)}\left(x^{k)}, y^{k)}; F\right)}{PBM^{b,k)}\left(x^{k)}; F\right) + PBM^{b,k)}\left(y^{k)}; F\right) - SBM^{b,k)}\left(x^{k)}, y^{k)}; F\right)}\right)$$

vale cero, entonces cada uno de los sumandos (al ser no negativos) ha de ser igual a cero y, por lo tanto, debido a la continuidad de F se obtiene que x tiene que ser igual a y. Es decir, que esta suma sobre las coordenadas es también una distancia. Por último, se tiene que $DBM(x, y; F, B)$ es una distancia debido a que es suma de distancias. ■

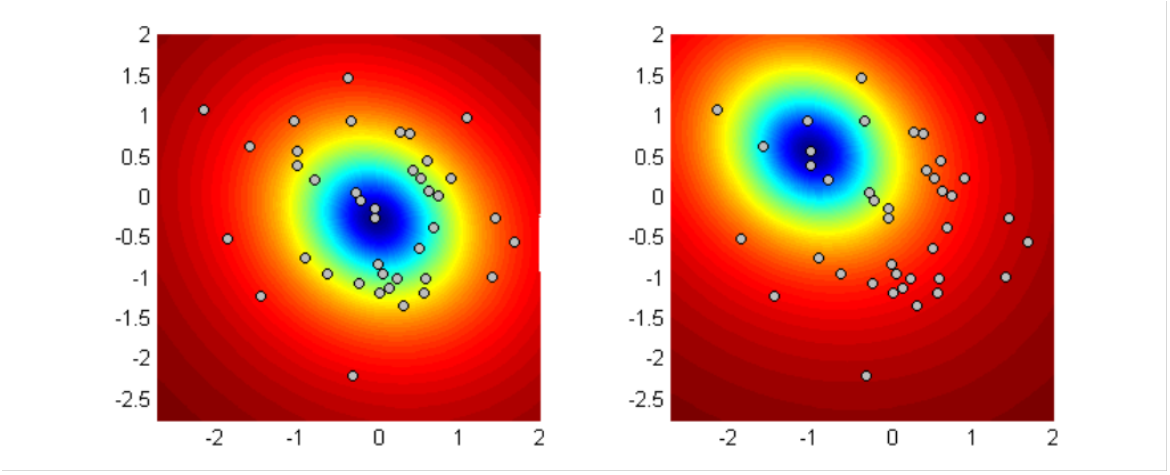
3.2. Algunos ejemplos prácticos

En este apartado se muestran los contornos de las funciones definidas en la sección anterior para los conjuntos de datos simulados según la normal y la exponencial, que se emplearon en el capítulo anterior.

Para las distancias basadas en profundidad construidas a partir de funciones de atipicidad o distancias, las ordernaciones y distancias que se obtienen son equivalentes a las que se obtienen a través de las similaridades por profundidad. A pesar de esta equivalencia, en esta sección se muestran las figuras para los dos conjuntos de datos. Las Figuras 3.2 y 3.3 se corresponden a las superficies para las muestras normal y exponencial respectivamente. Se oberva la similitud entre éstos y los presentados en el capítulo anterior. Lo mismo sucede para la distancia por proyecciones (Figuras 3.4 y 3.5). Aunque se ha probado que no es una distancia, se muestran también estas superficies para $1 - SO\left(x, y; F\right)$ (Figuras 3.6 y 3.7).

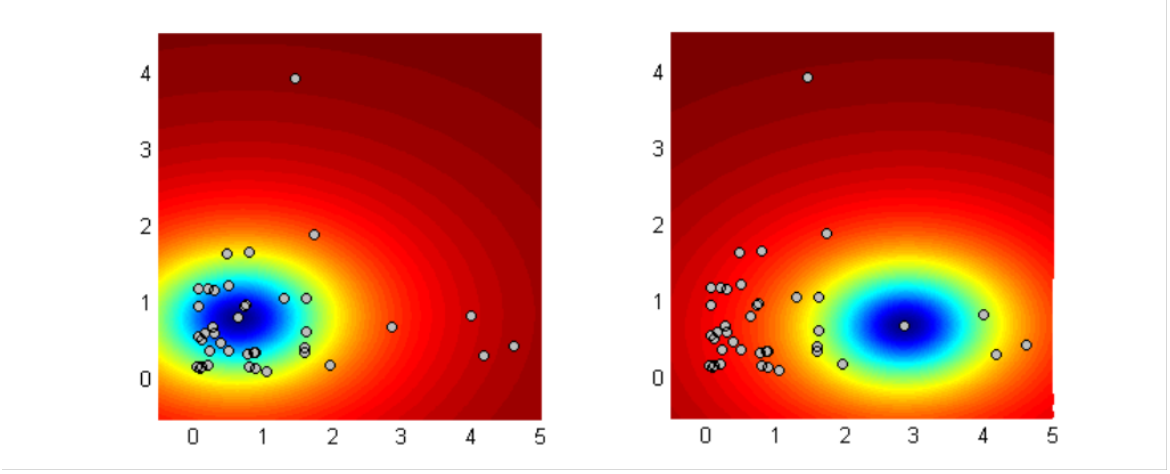
Para las otras tres distancias la situación es diferente debido a que la similaridad se estandariza por los valores de profundidad, es decir, que la distancia depende de lo cercanos al centro de la distribución que sean los puntos que se están comparando. Esto produce que los contornos de las distancias, fijado uno de los dos puntos, cambien sustancialmente cuando ese punto es próximo al centro, o no. En las figuras que se mostraron en el capítulo dos, el centro atraía los contornos hacia sí mismo, como un centro de gravedad. La situación después de la estandarización que produce la distancia es otra, el centro pierde peso y los contornos se aproximan al punto que se toma como fijo. Esos contornos parece que son más apropiados para medir distancias que los que se obtienen con las similaridades. Las Figuras 3.8 y 3.9 son las superficies de la distancia simplicial para los dos conjuntos de datos de la sección 2.3.

Las Figuras desde 3.10 hasta 3.13 son las correspondientes a las distancias por bandas y por bandas modificada. Se observa que los contornos se vuelven algo más simétricos con respecto al punto fijo, pero que aún así siguen respetando adecuadamente la forma de la nube de puntos.



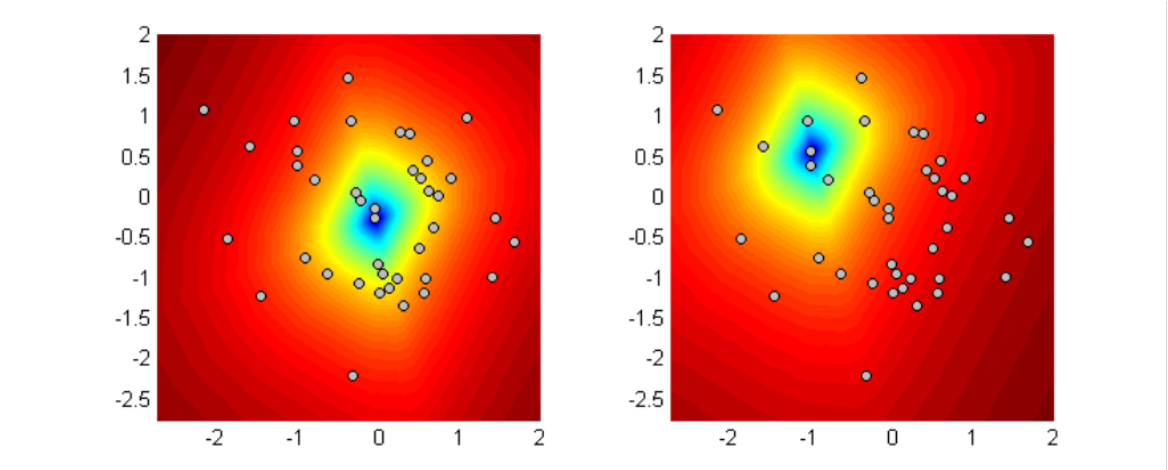
(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.2: *Distancia de Mahalanobis con respecto a un punto fijo para una muestra normal.*



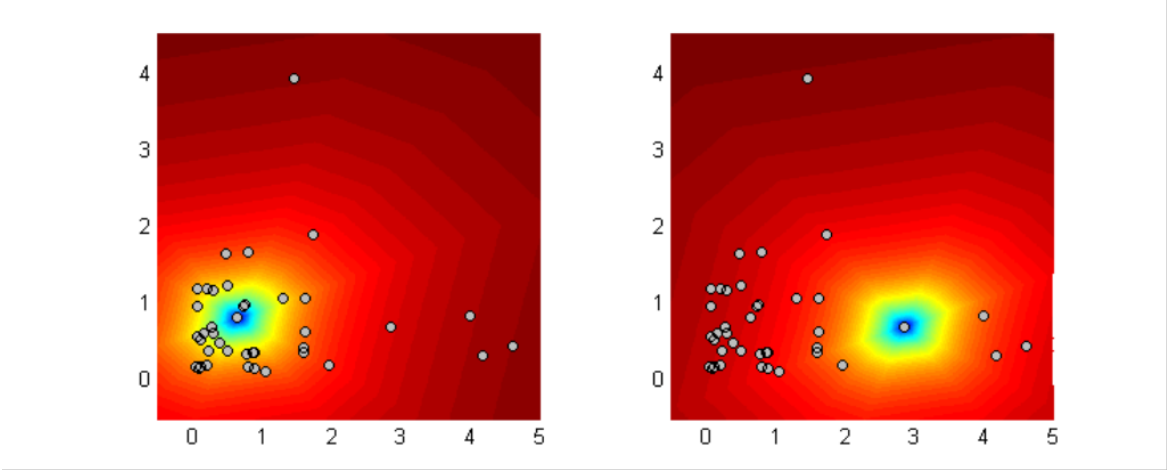
(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.3: *Distancia de Mahalanobis con respecto a un punto fijo para una muestra exponencial.*



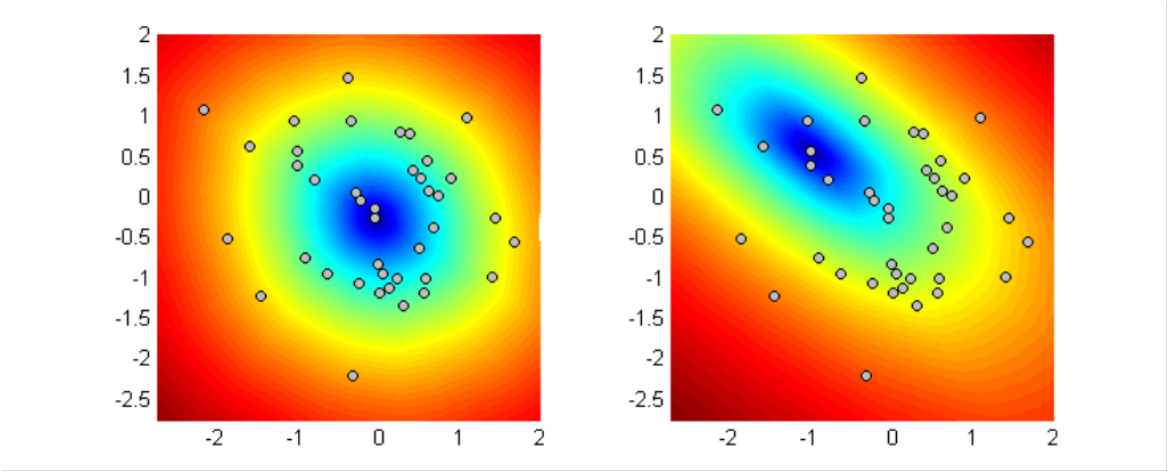
(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.4: *Distancia por proyecciones con respecto a un punto fijo para una muestra normal.*



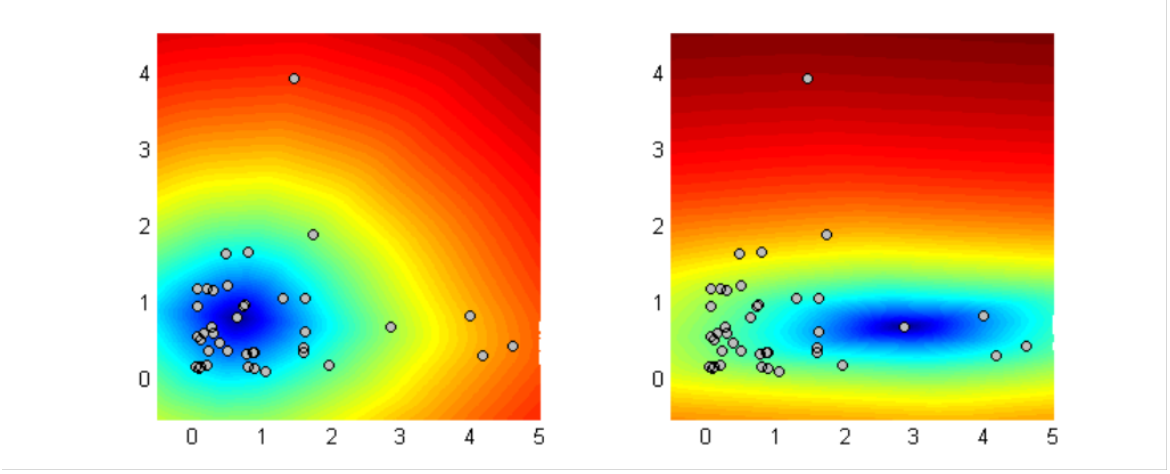
(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.5: *Distancia por proyecciones con respecto a un punto fijo para una muestra exponencial.*



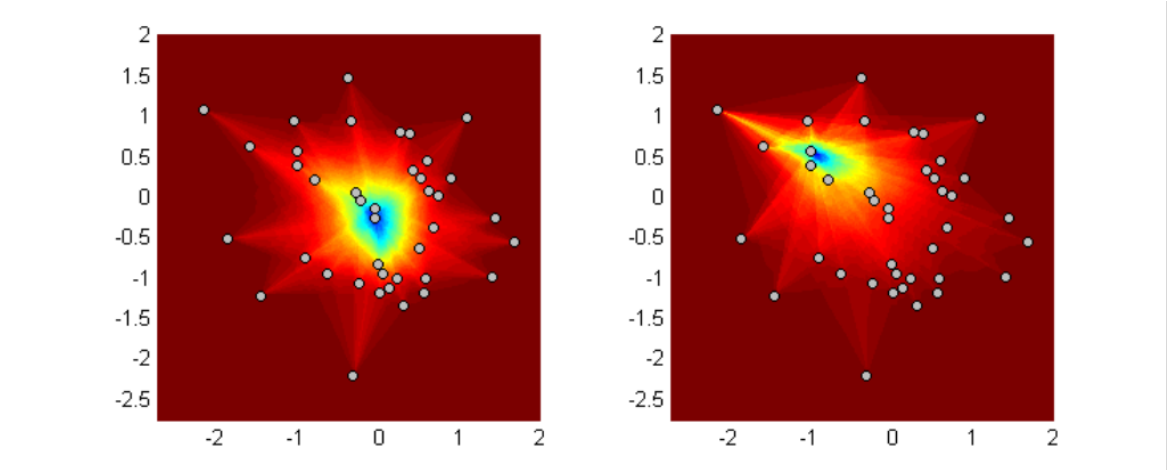
(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.6: *Uno menos la similaridad de Oja con respecto a un punto fijo para una muestra normal.*



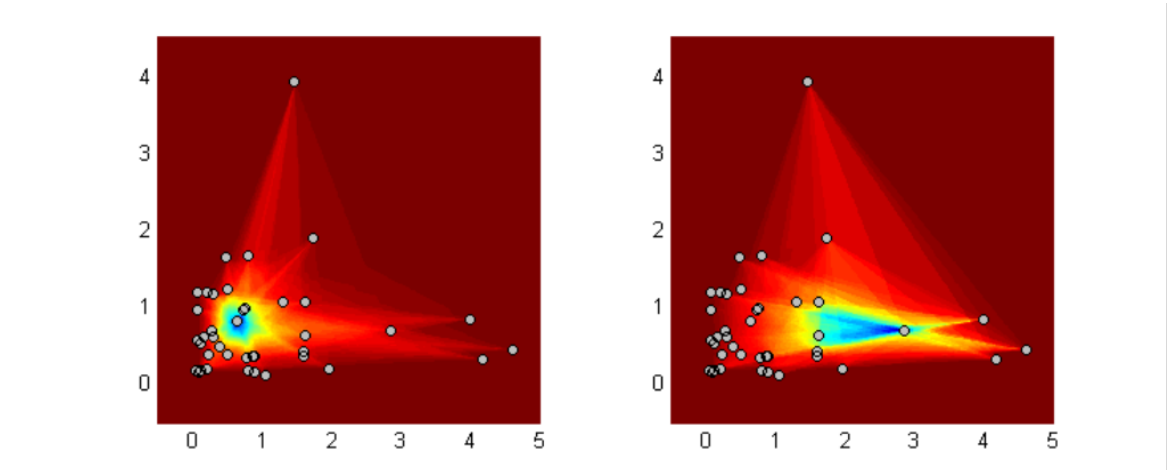
(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.7: *Uno menos la similaridad de Oja con respecto a un punto fijo para una muestra exponencial.*



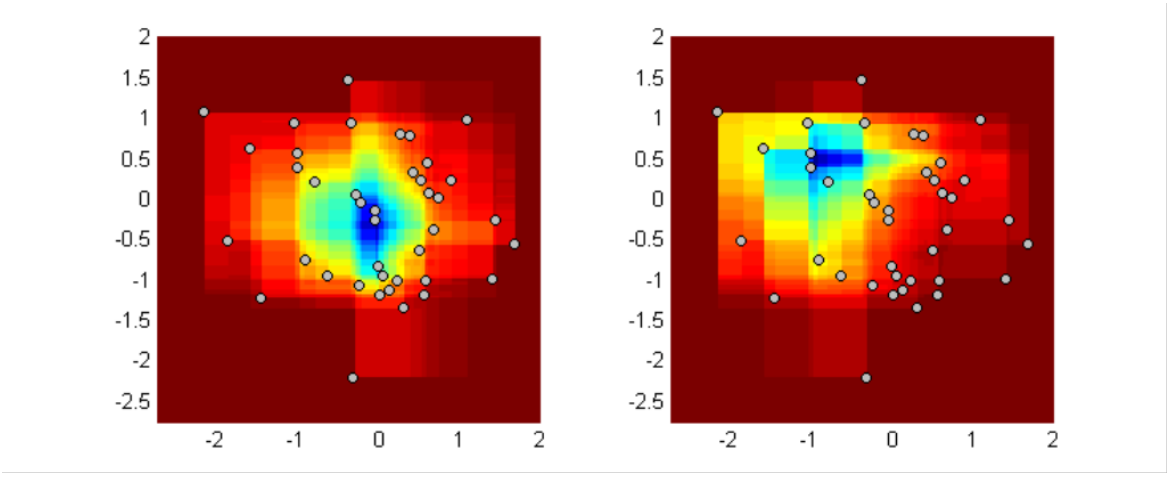
(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.8: *Distancia simplicial con respecto a un punto fijo para una muestra normal.*



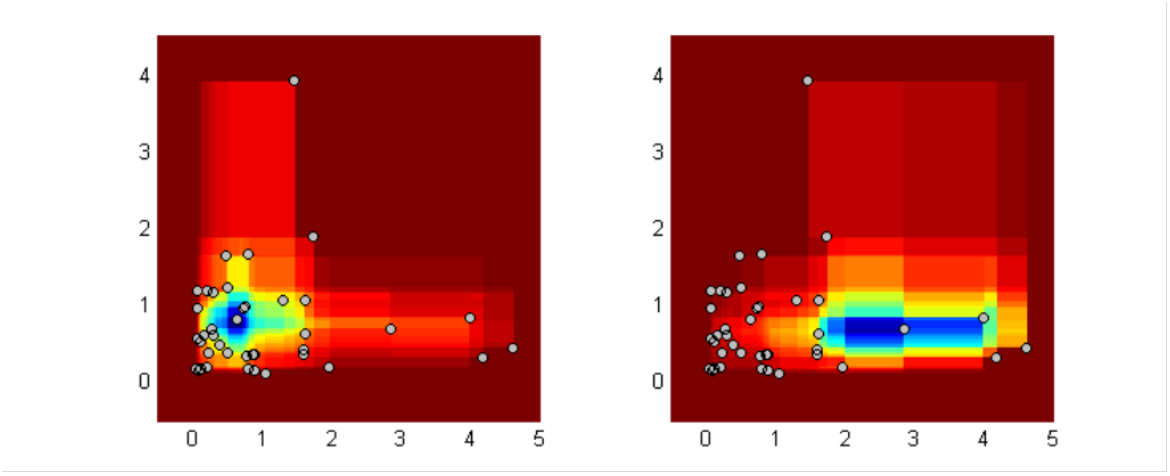
(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.9: *Distancia simplicial con respecto a un punto fijo para una muestra exponencial.*



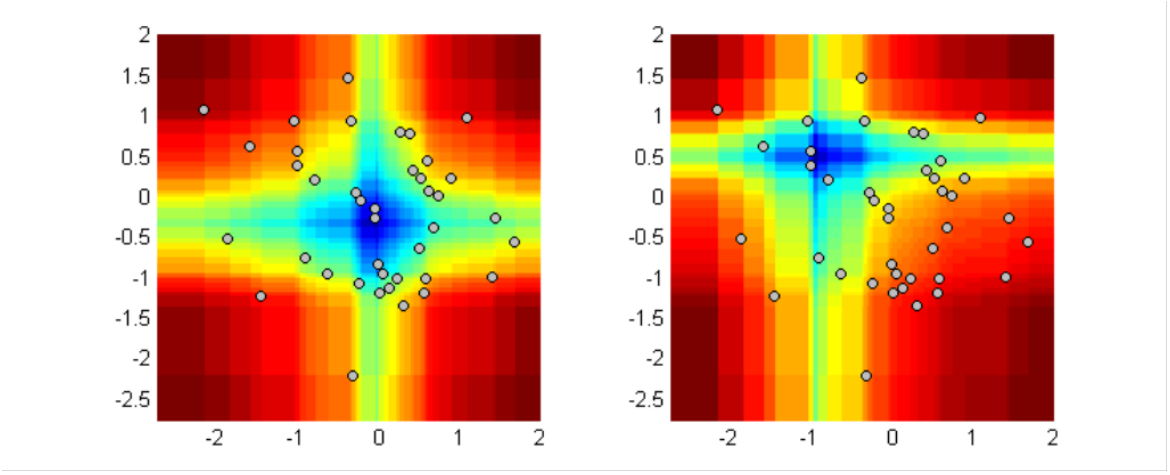
(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.10: *Distancia por bandas con respecto a un punto fijo para una muestra normal.*



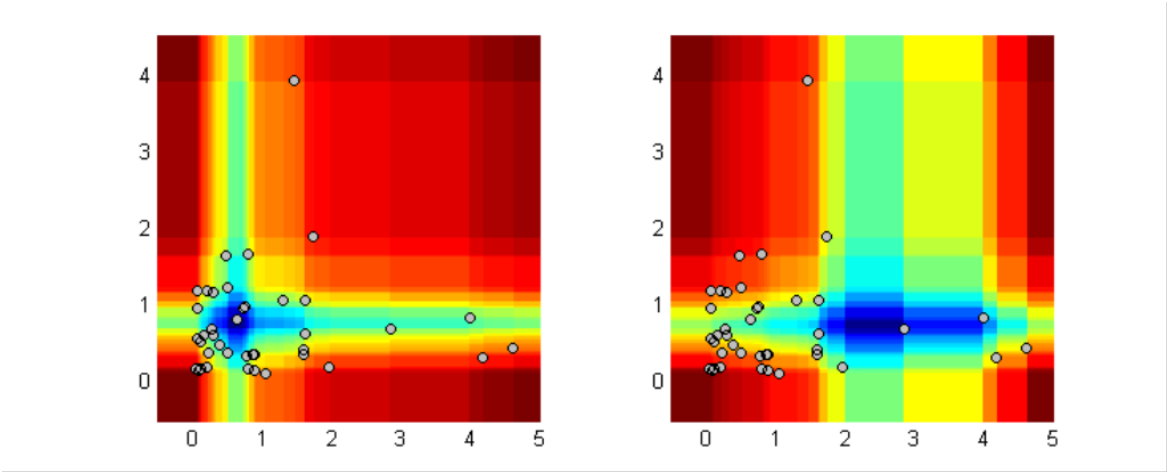
(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.11: *Distancia por bandas con respecto a un punto fijo para una muestra exponencial.*



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.12: *Distancia por bandas modificada con respecto a un punto fijo para una muestra normal.*



(a) Punto de referencia central. (b) Punto de referencia externo.

Figura 3.13: *Distancia por bandas modificada con respecto a un punto fijo para una muestra exponencial.*

3.3. Aplicación de las distancias basadas en profundidad

En esta sección se estudia el comportamiento de las distancias basadas en profundidad en el análisis de conglomerados. En primer lugar, se presenta una modificación del algoritmo de k-medias. Este algoritmo modificado, aplicado sobre las distancias por profundidad, mejora los resultados de agrupamiento que estas funciones obtendrían si se empleara el algoritmo de k-medias. En segundo lugar, se aplica dicho algoritmo modificado sobre conjuntos de datos bidimensionales. Éstos se simulan con los mismos modelos que se emplearon en el Capítulo 2. Por último, los resultados que se obtienen para las distancias por profundidad, se comparan, tanto con los obtenidos mediante k-medias (con la distancia euclídea), como con los obtenidos empleando la distancia euclídea en el algoritmo modificado.

Uno de los procedimientos de agrupamiento no jerárquico más empleado es el algoritmo de k-medias (Hartigan (1975)). Es un método iterativo que, partiendo de un conjunto de puntos considerados como los representantes de cada grupo y denominados centroides, asigna cada una de las observaciones al grupo de cuyo centroide está más próximo. Se recalculan los centroides y se reasignan las observaciones tantas veces como sean necesarias, hasta que los centroides se estabilizan y no se producen cambios en la asignación de observaciones. Se persigue por tanto, minimizar la suma de las distancias entre cada observación y el centroide de su grupo. Para poder emplear el algoritmo es necesario conocer el número de grupos en que se divide la muestra. En ocasiones, debido a la naturaleza de los datos, es posible tener esta información, pero en otras muchas situaciones esto no es posible y se hace necesario estimar dicho número. Para la elección de los centroides iniciales existen diversas propuestas que van desde asignar de forma aleatoria observaciones de la muestra, hasta ejecutar el algoritmo con submuestras del conjunto de datos y tomar los centroides de los grupos resultantes. El principal problema de este método es la sensibilidad a la elección del punto inicial, ya que el algoritmo puede detenerse en algún mínimo local que no sea del todo satisfactorio.

3.3.1. Modificación del algoritmo de k-medias

El algortimo de k-medias basa la asignación de observaciones en los centroides de cada grupo. La modificación que se introduce a este algoritmo es la eliminación del centroide y la asignación de observaciones al grupo para el que la distancia media sobre todos los elementos del grupo sea mínima. De este modo se consigue una asignación más robusta ya que no depende únicamente de la distancia a un sólo punto. La inicialización del algoritmo se conserva sin modificaciones, es decir, dados K puntos, donde K es el número de grupos, se asignan todas las observaciones al punto más cercano.

El algoritmo propuesto consta de los siguientes pasos:

1. Obtener los K representantes iniciales
2. Calcular la distancia de cada punto a los K puntos
3. Asignar cada observación al grupo con menor media de distancia a sus elementos
4. Repetir el paso 3 hasta que no se pueda minimizar más la suma de las distancias de las observaciones a su grupo

3.3.2. Análisis de sensibilidad de los puntos iniciales

En este apartado se analiza cómo influye la elección de los K puntos iniciales en el resultado final del porcentaje de observaciones bien agrupadas. Para poder obtener este porcentaje es necesario aplicar el algoritmo sobre conjuntos de observaciones simulados, para los que sí se conoce tanto la pertenencia a los grupos, como el número de éstos.

El algoritmo se aplica sobre los catorce conjuntos de datos empleados en el análisis de conglomerados jerárquico del Capítulo 2. Para cada uno de éstos se toman como puntos iniciales 1000 muestras sin reemplazamiento de tamaño el número de grupos y se ejecuta el algoritmo. Los porcentajes de acierto para cada ejemplo se representan mediante diagramas de caja que incluyen los resultados para las distancias por profundidad de Mahalanobis, proyecciones, simplicial, por bandas y por bandas modificada. A pesar de

que se ha demostrado que la transformación $1 - SO(x_i, x_j)$ no es una distancia, sus resultados también se incluyen. Por último, como distancia de contraste se presentan también los resultados obtenidos con el algoritmo modificado para la distancia euclídea (que se denota por DE) y los obtenidos con el algoritmo de k-medias y esta distancia.

Las matrices de distancias por profundidad empleadas en estas simulaciones se calculan con respecto a la mixtura al 50 % de la función empírica y la distribución normal de parámetros estimados, empleada en el análisis de conglomerados jerárquico del Capítulo 2.

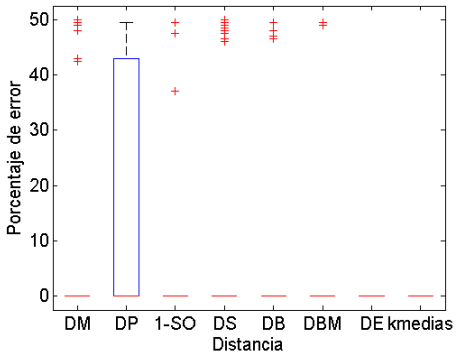
La organización de los resultados es análoga a la del capítulo anterior. En primer lugar, se muestran y analizan los resultados para los conjuntos de datos con grupos simétricos, seguidos de los asimétricos y de los que presentan relaciones no lineales. Tras el análisis individual de cada tipo de agrupación, se realiza un análisis conjunto para determinar qué similaridad presenta mejor comportamiento.

3.3.2.1. Grupos con distribución simétrica

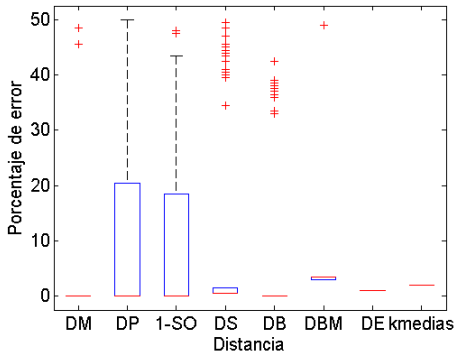
La Figura 3.14 contiene, para cada uno de los seis ejemplos de grupos simétricos, los diagramas de caja de los errores cometidos al aplicar el algoritmo modificado (además del de k-medias para la distancia euclídea) sobre las muestras analizadas en el Capítulo 2, tomando como representantes iniciales de cada grupo muestras aleatorias sin reemplazamiento del conjunto de datos. Se han simulado 1000 muestras de centros iniciales.

La sensibilidad a los puntos iniciales depende del rango de valores de los porcentajes de error. Aquéllas que tienen un rango más amplio son más sensibles a la elección de centros inicial y, por lo tanto, más necesaria es la aplicación de algún refinamiento en la selección de éstos. La valoración de los errores de clasificación como tal se realiza en la siguiente sección.

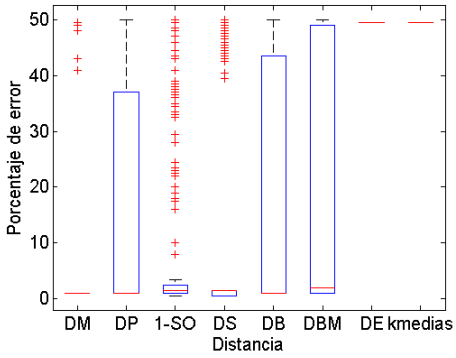
En cuanto a los resultados obtenidos para el ejemplo de los grupos de distribución normal estándar con vector de medias distinto (Figura 3.14(a)) la longitud de la caja es cero para todas las distancias, salvo para la de proyecciones. Todas las distancias propuestas en el capítulo presentan alguna observación atípica, lo que sugiere que, en



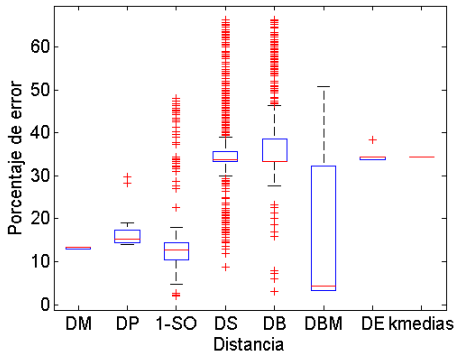
(a) Ejemplo 1



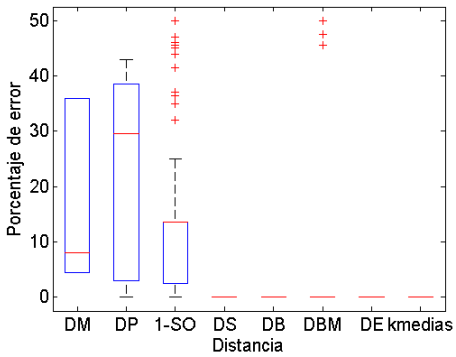
(b) Ejemplo 2



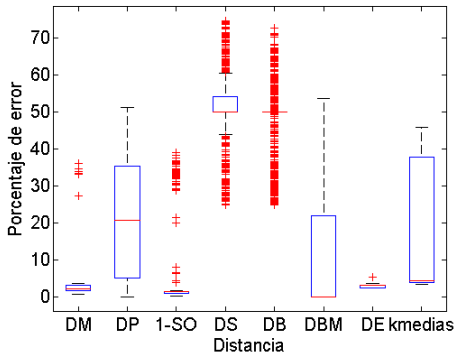
(c) Ejemplo 3



(d) Ejemplo 4



(e) Ejemplo 5



(f) Ejemplo 6

Figura 3.14: *Diagramas de caja del porcentaje de error para 1000 muestras de puntos iniciales en los ejemplos con grupos simétricos.*

esta situación, la distancia euclídea con el algoritmo modificado y con el de k-medias son las más robustas. Sobre el resto de los ejemplos se tiene que la distancia por proyecciones es la menos robusta, ya que en todos menos en un ejemplo, posee rangos intercuartílicos elevados, seguida de uno menos la similaridad de Oja y las distancias por bandas y por bandas modificadas que poseen rangos elevados o muchos valores atípicos en tres de los seis ejemplos. Del resto de funciones, la más robusta es la distancia euclídea con el algoritmo modificado.

3.3.2.2. Grupos con distribución asimétrica en al menos una coordenada

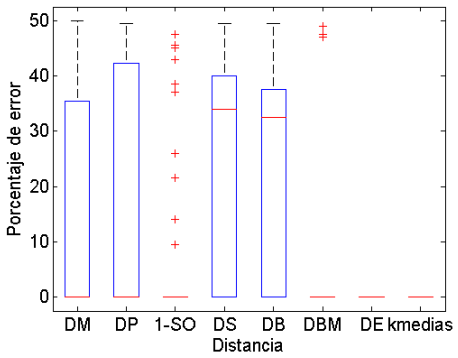
Cuando una de las coordenadas de las observaciones tiene una distribución asimétrica (Figura 3.15), la distancia por proyecciones continúa siendo la más sensible. En segundo lugar, se sitúan la similaridad de Oja y las distancias simplicial y de Mahalanobis, que se ve muy afectada por la asimetría de las variables. En una posición intermedia están las distancias por bandas y por bandas modificadas. Como en el caso anterior, las mejores son la distancia euclídea con el algoritmo modificado y el algoritmo de k-medias, que mejora con respecto al caso anterior.

3.3.2.3. Grupos con relaciones no lineales entre variables

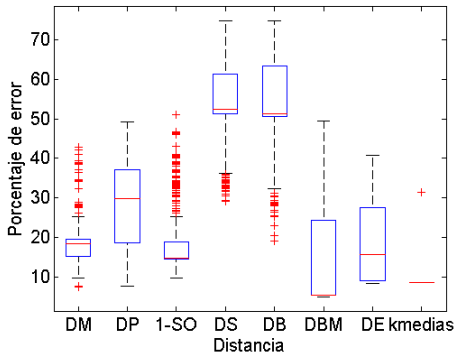
Por último, se analiza el tipo de muestras con grupos no lineales (Figura 3.16), teniéndose resultados análogos a los dos casos anteriores. Para las distancias basadas en profundidad se produce un empeoramiento generalizado y en especial, para la de Mahalanobis, que ahora se muestra como la menos robusta junto con la de proyecciones. Para el algoritmo modificado con la distancia euclídea, se tiene que, aunque el rango intercuartílico no aumenta considerablemente, aparece un mayor número de observaciones atípicas.

3.3.2.4. Comparación global

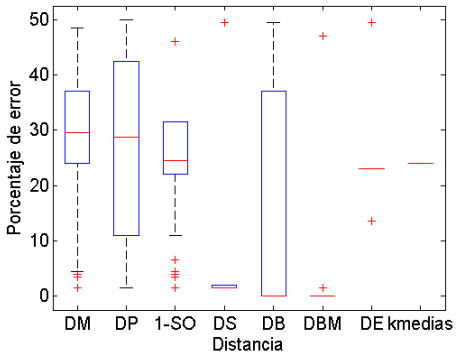
Finalmente, se comparan todas las distancias y algoritmos por medio de los rangos intercuartílicos y la proporción de atípicos de cada ejemplo (Tablas 3.1 y 3.2).



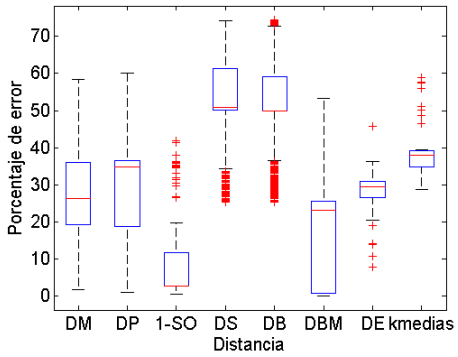
(a) Ejemplo 1



(b) Ejemplo 2



(c) Ejemplo 3

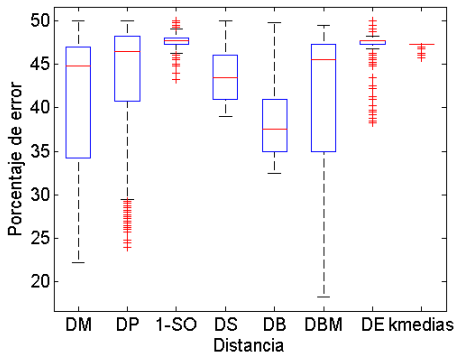


(d) Ejemplo 4

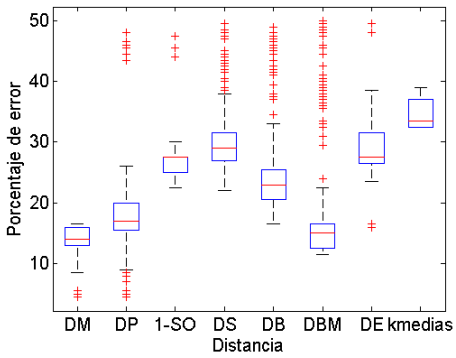
Figura 3.15: *Diagramas de caja del porcentaje de error para 1000 muestras de puntos iniciales en los ejemplos con grupos asimétricos.*

El algoritmo modificado con la distancia euclídea mejora ligeramente la dependencia de los centros iniciales del algoritmo de k-medias, que es la segunda de las opciones más robusta (véase la Tabla 3.1). De las funciones basadas en profundidad las que, en media, tienen mayor robustez son: uno menos la similaridad de Oja y la distancia simplicial, que poseen rangos medios muy similares. Las siguientes más sensibles son: la distancia por profundidad de Mahalanobis y las de bandas y bandas modificada con rangos entre 12 y 14 por ciento. Por último, se encuentra la distancia por proyecciones, cuyo rango medio se eleva hasta casi el 25 por ciento.

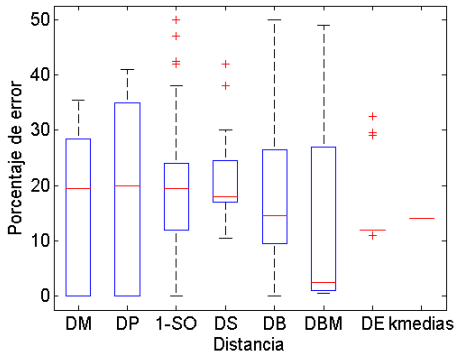
En los diagramas de caja anteriores se ha representado el resultado de la ejecución del método de agrupación, es decir, los porcentajes de error son los obtenidos para los



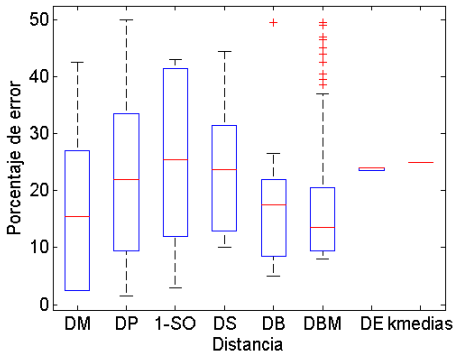
(a) Ejemplo 1



(b) Ejemplo 2



(c) Ejemplo 3



(d) Ejemplo 4

Figura 3.16: *Diagramas de caja del porcentaje de error para 1000 muestras de puntos iniciales en los ejemplos con grupos circulares y grupos con formas no lineales.*

mínimos (locales o globales) en los que el algoritmo se ha detenido. Por eso, debido a que a cada mínimo se puede llegar desde distintos puntos iniciales, es necesario calcular el porcentaje de atípicos para obtener una idea correcta de los comportamientos extremos de los mínimos alcanzados. Es decir, puede que el valor atípico representado en los diagramas esté repetido un elevado número de veces y, sin embargo, se tenga la impresión de que es una única solución la que tuvo ese comportamiento. Estos porcentajes de atípicos se encuentran recogidos en la Tabla 3.2. Según los datos de esta tabla, la que menor porcentaje de atípicos tiene es la distancia por proyecciones, seguida de la distancia por bandas modificada y de la distancia euclídea para los dos algoritmos. Aunque para interpretar estos porcentajes medios, es necesario tener en cuenta los valores de los rangos

		Distancia Euclídea		Similaridades					
Tipo	Ejemplo	k-medias mod.	k-medias	DM	DP	1-SO	DS	DB	DBM
Simétricos	1	0.0	0.0	0.0	43.0	0.0	0.0	0.0	0.0
	2	0.0	0.0	0.0	20.5	18.5	1.0	0.0	0.5
	3	0.0	0.0	0.0	36.0	1.5	1.0	42.5	48.0
	4	0.7	0.0	0.3	3.0	4.0	2.3	5.3	29.0
	5	0.0	0.0	31.5	35.5	11.0	0.0	0.0	0.0
	6	0.8	33.8	1.5	30.3	0.5	4.3	0.0	22.0
Asimétricos	1	0.0	0.0	35.5	42.3	0.0	40.0	37.5	0.0
	2	18.3	0.0	4.3	18.4	4.5	10.1	12.8	19.0
	3	0.0	0.0	13.0	31.5	9.5	0.5	37.0	0.0
	4	4.5	4.5	16.8	17.8	9.0	11.0	9.1	24.8
No lineales	1	0.5	0.0	12.8	7.5	0.8	5.0	6.0	12.3
	2	5.0	4.5	3.0	4.5	2.5	4.5	5.0	4.0
	3	0.0	0.0	28.5	35.0	12.0	7.5	17.0	26.0
	4	0.5	0.0	24.5	24.0	29.5	18.5	13.5	11.0
	Media	2.2	3.1	12.3	24.9	7.4	7.6	13.3	14.0

Tabla 3.1: *Rango intercuartílico de los errores de agrupación.*

intercuartílicos, ya que, para una distancia con valores altos del rango, es menos probable que aparezcan observaciones atípicas debido a que las cantidades que se analizan son porcentajes. Es lo que sucede con la distancia por proyecciones que presenta los rangos más elevados y el porcentaje de atípicos más pequeño. Sin embargo, sí permite comparar las distancias que presentan rangos parecidos. Por ejemplo, para el algoritmo k-medias el porcentaje de datos extremos es menos de la mitad que para la distancia euclídea para el algoritmo modificado.

Como conclusión final se tiene que, debido a la falta de robustez frente a la inicialización del algoritmo y debido a que tiene más mínimos locales que k-medias, el uso de las distancias basadas en profundidad debe estar acompañado de algún tipo de refinamiento de los puntos iniciales o de algún criterio que minimice su impacto.

		Distancia Euclídea		Similaridades					
Tipo	Ejemplo	k-medias mod.	k-medias	DM	DP	1-SO	DS	DB	DBM
Simétricos	1	0.0	0.0	9.2	0.0	1.4	1.5	1.9	0.3
	2	0.0	0.0	0.3	0.0	0.2	2.9	15.2	0.1
	3	0.0	0.0	9.1	0.0	7.3	8.8	0.0	0.0
	4	0.3	0.0	0.0	0.2	5.7	26.5	23.8	0.0
	5	0.0	0.0	0.0	0.0	19.9	0.0	0.0	0.3
	6	0.1	0.0	11.0	0.0	22.6	24.4	38.3	0.0
Asimétricos	1	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.4
	2	0.0	14.5	22.7	0.0	18.4	3.0	1.8	0.0
	3	0.3	0.0	7.9	0.0	8.7	0.3	0.0	0.2
	4	11.5	3.1	0.0	0.0	2.4	5.3	15.2	0.0
No lineales	1	22.3	8.5	0.0	13.7	19.7	0.0	0.0	0.0
	2	0.7	0.0	3.8	8.1	0.5	5.8	6.8	21.9
	3	23.7	0.0	0.0	0.0	0.7	11.7	0.0	0.0
	4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.1
	Media	4.3	1.9	4.6	1.6	7.8	6.4	7.4	1.7

Tabla 3.2: *Porcentaje de atípicos en los diagramas de caja.*

3.3.3. Resultados de simulación

En el apartado anterior no se han analizado los errores de clasificación desde el punto de vista de la bondad del agrupamiento. Ese análisis se lleva a cabo a continuación, por medio de simulaciones de 100 muestras para cada uno de los modelos con los que se generaron las muestras del apartado anterior. Además, con el objetivo de evitar la falta de robustez de las distancias por profundidad ya comentada, para cada una de las 100 muestras de cada ejemplo, se ha ejecutado el algoritmo en 20 ocasiones (20 conjuntos de centros iniciales, seleccionados sin reeemplazamiento de la muestra) y se ha tomado como mejor solución de cada muestra la repetición para la que las distancias intra grupos era menor.

3.3.3.1. Grupos con distribución simétrica

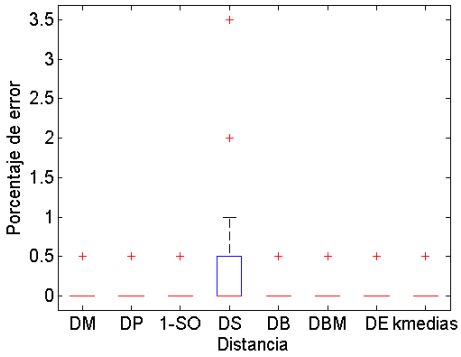
La Figura 3.17 contiene los diagramas de caja para los mejores resultados de las 100 muestras generadas con el modelo de cada uno de los ejemplos de grupos simétricos. Para el ejemplo uno, con grupos de distribución normal estándar y de medias distintas, todas las distancias y algoritmos, salvo la simiplicial, presentan errores menores al 1 %, aunque mayoritariamente no se cometen errores. Cuando los grupos presentan correlación de sentido inverso y medias distintas (Figura 3.17(b)), todas las distancias por profundidad, salvo la de bandas modificada, presentan errores de clasificación menores que los de k-medias y del algoritmo modificado usando la distancia euclídea.

Cuando la variabilidad de una de las coordenadas es superior a la de la otra (Figura 3.17(c)), la distancia euclídea comete errores cercanos al 45 %, mientras que las distancias por profundidad el error se sitúa en torno al 1 %. Si se incluye además otro grupo con distribución normal estándar y media distinta de la de los anteriores (Figura 3.17(d)), la distancia euclídea disminuye ligeramente su error, mientras que la distancia por profundidad de Mahalanobis, por proyecciones y uno menos la similaridad de Oja aumentan el error drásticamente hasta el entorno del 50 %. El resto de distancias tiene errores en el entorno del 5 % o menos, siendo la de bandas modificada la que menores y menos variables errores tiene.

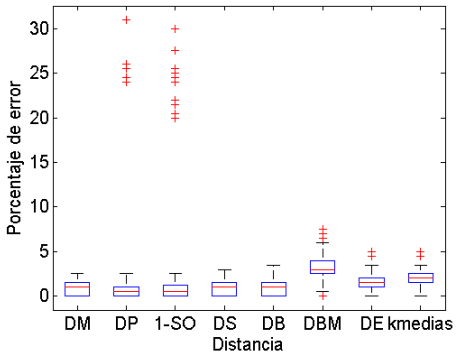
Por último, para las muestras con grupos rectangulares, los peores errores se obtienen para la distancia por profundidad de Mahalanobis, la de proyecciones y uno menos la similaridad de Oja. Los resultados para la distancia euclídea son ligeramente peores que para las distancias simplicial, por bandas y por bandas modificada, cuando hay dos grupos. Para las muestras de cuatro grupos la distancias de los errores entre simplicial, bandas y bandas modificadas y el resto aumentan.

3.3.3.2. Grupos con distribución asimétrica en al menos una coordenada

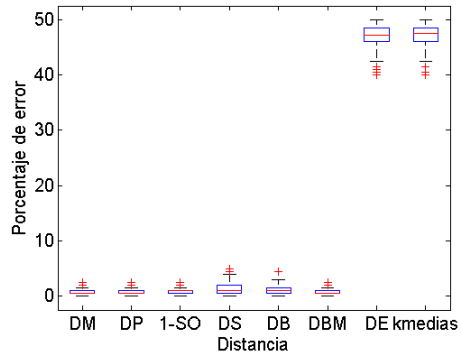
Para los modelos en que los que alguna de las coordenadas se distribuye según una exponencial, los porcentajes de error de clasificación más elevados, corresponden a la distancia por profundidad de Mahalanobis, a la de proyecciones y a uno menos la similaridad



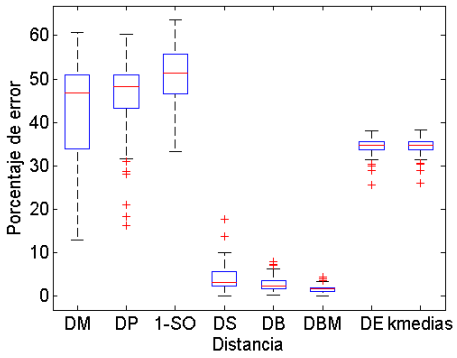
(a) Ejemplo 1



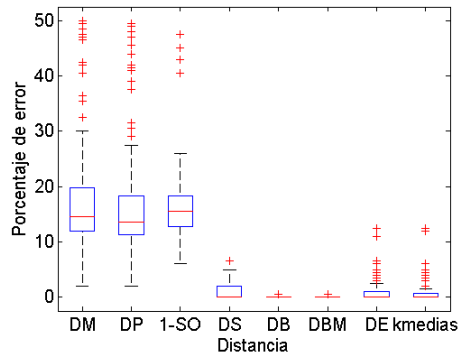
(b) Ejemplo 2



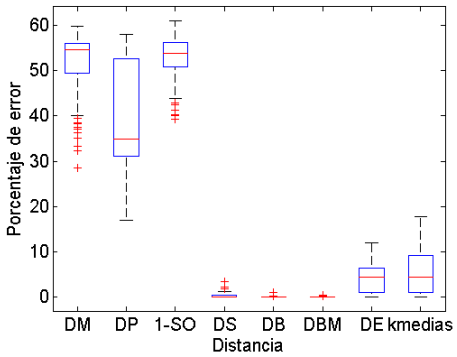
(c) Ejemplo 3



(d) Ejemplo 4



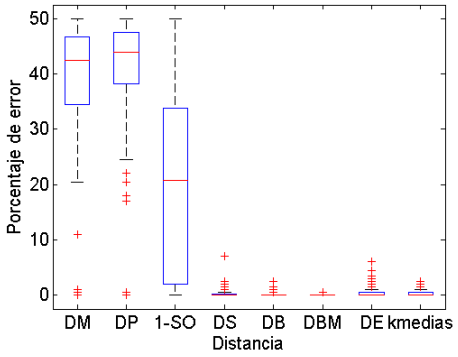
(e) Ejemplo 5



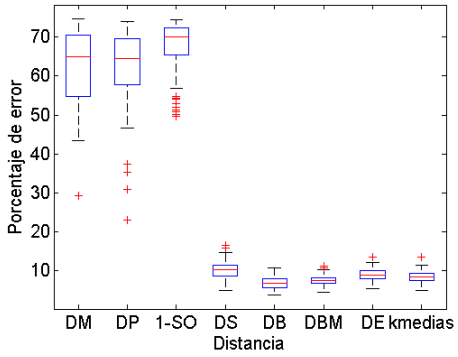
(f) Ejemplo 6

Figura 3.17: *Diagramas de caja del porcentaje de error para la mejor agrupación, para 100 muestras de los ejemplos con grupos simétricos.*

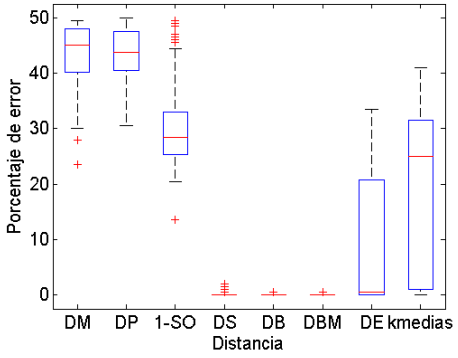
de Oja. Para el resto de distancias hay que diferenciar los casos en que las dos coordenadas son exponenciales de los que tienen forma rectangular (coordenada uniforme y coordenada exponencial). En el primer grupo (Figuras 3.18(a) y 3.18(b)) todas tienen resultados similares, si bien, tanto la distancia por bandas como por bandas modificada presentan resultados mejores que las otras tres. En el segundo grupo (Figuras 3.18(c) y 3.18(d)), los errores para la distancia euclídea aumentan de forma considerable, especialmente cuando son cuatro los grupos.



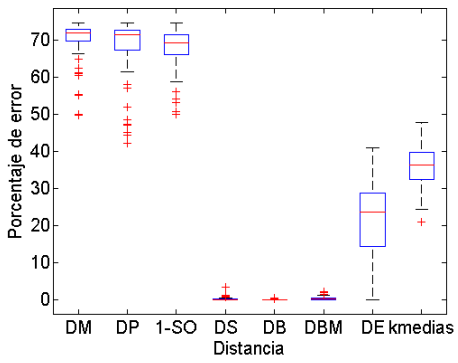
(a) Ejemplo 1



(b) Ejemplo 2



(c) Ejemplo 3

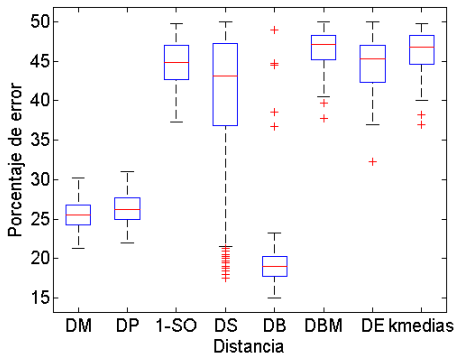


(d) Ejemplo 4

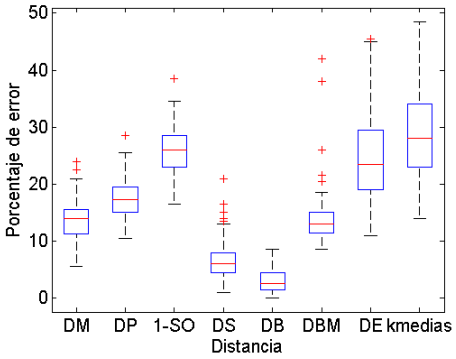
Figura 3.18: *Diagramas de caja del porcentaje de error para la mejor agrupación, para 100 muestras de los ejemplos con grupos asimétricos.*

3.3.3.3. Grupos con relaciones no lineales entre variables

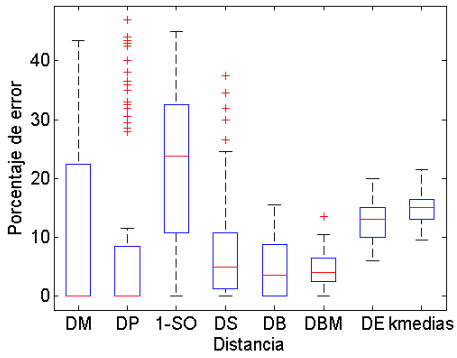
Para finalizar los análisis de resultados individuales, se estudia el error de agrupamiento para las muestras con grupos no lineales. En el primero de los modelos (circunferencia y anillo, Figura 3.19(a)), se observa un incremento generalizado del error. Las distancias que mejor comportamiento tenían en los ejemplos anteriores son las que ahora cometen más error. La que presenta mejores porcentajes es la distancia simplicial.



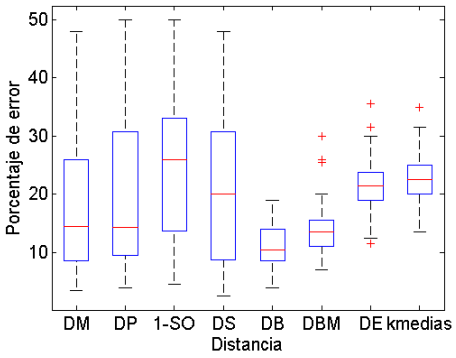
(a) Ejemplo 1



(b) Ejemplo 2



(c) Ejemplo 3



(d) Ejemplo 4

Figura 3.19: *Diagramas de de caja del porcentaje de error para la mejor agrupación, para 100 muestras de los ejemplos con grupos con formas no lineales.*

En el segundo (circunferencia y mitad de anillo, Figura 3.19(b)), las similaridades que menores errores cometen vuelven a ser la de bandas, simplicial y bandas modificada, que se sitúan muy alejadas de la distancia euclídea con el algoritmo modificado y con k-medias.

En el tercer ejemplo (Figura 3.19(c)), los mayores errores se obtienen para la distancia euclídea y para uno menos la similaridad de Oja. Mientras que en el último (Figura 3.19(d)), los errores más elevados se producen tanto para la distancia euclídea como para la distancia simplicial y la similaridad de Oja.

3.3.3.4. Comparación global

De forma global (Tabla 3.3) se confirman los resultados que se han ido obteniendo por grupos, es decir, que las distancias que ofrecen mejores agrupamientos son la de bandas, seguida de la simplicial y de la de bandas modificada, con 3.3, 6.4 y 6.5 por ciento de error, respectivamente. Además, mejoran sustancialmente los resultados obtenidos tanto por el algoritmo de k-medias, como el algoritmo modificado, para la distancia euclídea.

		Distancia Euclídea		Similaridades					
Tipo	Ejemplo	k-medias mod.	k-medias	DM	DP	1-SO	DS	DB	DBM
Simétricos	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	2	1.5	2.0	1.0	0.5	0.5	1.0	1.0	3.0
	3	47.3	47.5	0.5	0.5	1.0	1.0	1.0	0.5
	4	34.7	34.7	46.8	48.2	51.3	3.2	2.3	1.7
	5	0.0	0.0	14.5	13.5	15.5	0.0	0.0	0.0
	6	4.5	4.5	54.6	34.9	53.8	0.0	0.0	0.0
Asimétricos	1	0.0	0.0	42.5	44	20.8	0.0	0.0	0.0
	2	9.0	8.5	65.0	64.6	70.1	10.3	6.9	7.5
	3	0.5	25.0	45.0	43.8	28.5	0.0	0.0	0.0
	4	23.8	36.4	72.0	71.5	69.4	0.0	0.0	0.3
No lineales	1	45.3	46.8	25.5	26.3	44.9	43.1	19.0	47.1
	2	23.5	28.0	14.0	17.3	26.0	6.0	2.5	13.0
	3	13.0	15.0	0.0	0.0	23.8	5.0	3.5	4.0
	4	21.5	22.5	14.5	14.3	26.0	20.0	10.5	13.5
	Media	16.0	19.4	28.3	27.1	30.8	6.4	3.3	6.5

Tabla 3.3: *Mediana del porcentaje del error.*

En cuanto al porcentaje de atípicos presentes en las muestras del porcentaje de error

(Tabla 3.4), se tiene en todos los casos valores entre el 3 y el 6 por ciento, lo que sugiere que tomando la solución de menor distancia para los 20 conjuntos de puntos iniciales se ha conseguido homogeneizar la dependencia de las distancias y los centros iniciales, obteniéndose, por lo tanto, que los resultados entre las diferentes distancias sean comparables.

		Distancia Euclídea		Similaridades					
Tipo	Ejemplo	k-medias mod.	k-medias	DM	DP	1-SO	DS	DB	DBM
Simétricos	1	4	4	3	3	3	3	3	3
	2	3	4	0	6	13	0	0	5
	3	5	5	5	6	8	3	2	7
	4	6	6	0	6	0	2	3	4
	5	14	16	18	16	4	1	4	2
	6	0	0	12	0	6	3	12	6
Asimétricos	1	15	9	7	6	0	9	12	1
	2	1	1	1	4	9	2	0	2
	3	0	0	2	0	10	16	2	2
	4	0	1	9	10	6	10	23	4
No lineales	1	1	2	0	0	0	16	6	3
	2	1	0	2	1	1	5	0	5
	3	0	0	0	22	0	5	0	1
	4	3	1	0	0	0	0	0	3
	Media	3.8	3.5	4.2	5.7	4.3	5.4	4.8	3.4

Tabla 3.4: *Porcentaje de atípicos en los diagramas de caja.*

Capítulo 4

Contrastes basados en profundidad

Resumen

Este capítulo se centra en contrastes basados en profundidad para espacios de dimensión mayor que uno. En él se introducen tres métodos para medir la discrepancia entre muestra y población. El primero se basa en la dispersión muestral representada a través de la curva de escala introducida en el primer capítulo. En segundo lugar, también basado en curvas, se presenta un contraste en el que el estadístico de discrepancia se obtiene a través de comparaciones de las regiones centrales de muestra y población. Por último, se define un contraste basado en las similaridades definidas en el Capítulo 2. Para los tres contrastes que se proponen se realiza un estudio de potencia mediante simulación, cuyos resultados se comparan entre sí y con los contrastes de bondad de ajuste multivariantes más relevantes de la literatura. Se estudia la potencia para las distribuciones normal, uniforme y exponencial. Para las tres distribuciones y de forma global sobre los tres contrastes, las profundidades y similaridades de Oja y de bandas modificada, son las que obtienen los mejores resultados. En particular, para la distribución normal multivariante, el contraste basado en la similaridad de Oja y el basado en la similaridad por bandas modificada, se comportan de forma muy competitiva con respecto a otros contrastes de bondad de ajuste, mejorando los porcentajes de rechazo de muchos de éstos.

4.1. Introducción

Uno de los problemas básicos de la estadística aplicada consiste en la elección del modelo probabilístico para un conjunto de observaciones. Este problema se hace más relevante cuando se realiza la diagnosis de un modelo basado en hipótesis distribucionales de la muestra o de los residuos de dichos modelos. Teniendo que, de no cumplirse las hipótesis, el modelo y sus propiedades pueden no ser válidos. De ahí la importancia de contrastar los supuestos iniciales y de que estos contrastes sean lo más potentes posible.

Dada una muestra aleatoria simple $x_1, x_2, \ldots, x_n$ cuya función de distribución F es desconocida, los tests de bondad de ajuste contrastan las siguientes hipótesis:

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0,$$

donde F_0 es la función de distribución que se supone ha generado dicha muestra.

Existe una gran cantidad de contrastes en la literatura para el caso univariante. Algunos pueden aplicarse sobre cualquier función de distribución, mientras que otros son específicos para determinadas funciones como, por ejemplo, la distribución normal que es la que, por su importancia, ha sido estudiada de una forma más amplia. Algunos ejemplos de contrastes univariantes son el de Anderson y Darling (1954), el de Kolmogorov-Smirnov y Cramer-von Mises (Darling (1957)), el contraste de la chi-cuadrado (Watson (1957), Watson (1958) y Watson (1959)), el contraste de análisis de varianza para normalidad (Shapiro y Wilk (1965)) y el de D'Agostino (1971). En Stephens (1974) se propone un contraste basado en la función de distribución empírica, en Royston (1982b) y Royston (1982a) se estudia y extiende el estadístico de Shapiro-Wilk para la distribución normal. Contrastes basados en la función característica pueden encontrarse en Epps y Pulley (1983) y Hall y Welsh (1983); basado en la idea de entropía está Vasicek (1976). Una modificación del contraste de Anderson y Darling fue propuesta por Sinclair et al. (1990). En Csörgo y Faraway (1996) se realiza un estudio de las distribuciones asintóticas del estadístico de Cramer-von Mises. Estudios en los que se realizan comparaciones de contrastes pueden verse en Shapiro et al. (1968), Stephens (1974), Pearson et al. (1977) y Baringhaus et al. (1989).

El caso multidimensional no ha sido tan analizado como el unidimensional. Entre los contrastes más importantes se encuentran los basados en la asimetría y la curtosis multivariante (Mardia (1970) y Malkovic y Afifi (1973)), en el estadístico de Shapiro-Wilk (Royston (1983), Fattorini (1986) y Koziol (1986)) y en el estadístico Cramer-von Mises (Zhu et al. (1997)). La extensión al caso multivariante del contraste de Kolmogorov-Smirnov puede encontrarse en Justel et al. (1997). Existen propuestas en las que se realizan proyecciones (Cui y Cheng (1996) y Zhu et al. (1997)), en las que se emplea la idea de entropía (Zhu et al. (1995)), las distancias entre puntos (Bartoszynski et al. (1997) y Székely y Rizzo (2005)) o se utiliza una metodología de vecinos más próximos (Zhou y Jammalamadaka (1993)). En Henze y Zirkler (1990), Romeu y Ozturk (1993) y Quiroz y Dudley (1991) pueden encontrarse otras propuestas.

Aunque los contrastes que se definen a continuación pueden aplicarse sobre cualquier distribución continua, la motivación de los mismos se centran en la distribución normal multivariante. Los estudios de potencia se realizan para la distribución normal multivariante y para dos vectores aleatorios bidimensionales, cuyas coordenadas son independientes y se distribuyen según una uniforme o una exponencial.

4.2. Contraste de dispersión basado en la curva de escala

El primer contraste que se propone se basa en una característica de forma de los datos: la dispersión medida a través de la curva de escala. A diferencia de los contrastes en Jarque y Bera (1987) y Mardia (1970), que se construyen a partir de la asimetría y curtosis para toda la muestra, el contraste que se introduce a continuación está definido sobre una curva, con lo que se dispone de toda la evolución de la característica de forma elegida. El patrón con que la versión muestral de dicha curva crece entre su valor mínimo (cero) y su valor máximo (el volumen de la envolvente convexa de todos los puntos) presenta comportamientos equivalentes para algunas funciones de distribución entre las que se encuentra la distribución sobre la que se realizan las simulaciones: la normal. Una desviación

elevada entre la curva muestral y la que se obtendría con la función de distribución que se desea contrastar sugiere que la muestra no procede de dicha distribución.

La curva de escala introducida en el Capítulo 1 calcula el volumen de las regiones centrales p-ésimas o C_p (véase la definición 1.4) para valores de p entre cero y uno.

Definición 4.1 *Sea F una función de distribución en $\mathbb{R}^d$ y $P(x;F)$ una función de profundidad con respecto a F. Se define la curva de escala como*

$$S(p) = \mathit{Volumen}(C_p),\ p \in [0,1],$$

donde $C_p = \underset{t}{\cap}\{R(t) : \mathit{Probabilidad}(R(t)) \geq p\}$ *y* $R(t) = \left\{x \in \mathbb{R}^d : P(x;F) > t\right\}$.

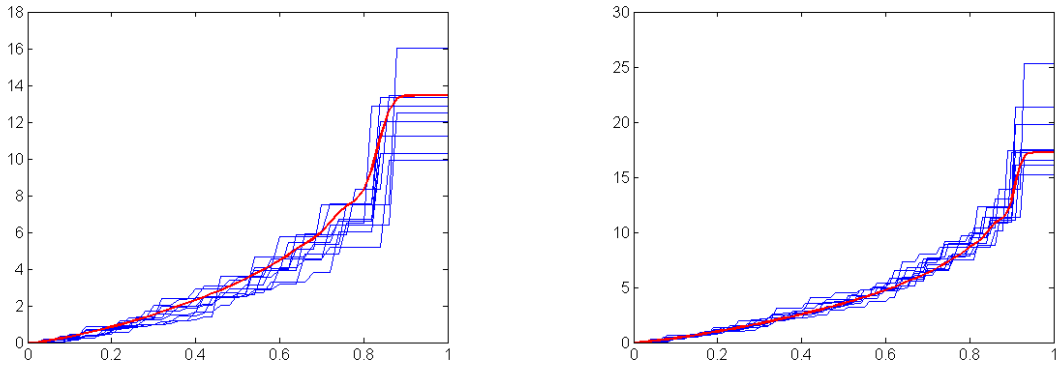
Dada una muestra $x_1, x_2, \ldots, x_n$ en $\mathbb{R}^d$, la curva de escala muestral o $S_n(p)$ se obtiene a partir de una estimación de las regiones centrales $C_{n,p}$. Una posible estimación de éstas consiste en tomar la envolvente convexa de los $\lceil np \rceil$ puntos más profundos, donde $\lceil np \rceil$ es igual a np, si np es entero, y a la parte entera de np más uno, si np no es entero. Es decir, dada la muestra con los puntos ordenados según alguna función de profundidad de más a menos profundos, $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$, se calcula

$$S_{n,p} = \text{Volumen}\left(\text{envolvente convexa}\left\{x_{[1]}, x_{[2]}, \ldots, x_{[\lceil np \rceil]}\right\}\right).$$

Para motivar el contraste se introducen a continuación varios ejemplos que muestran el comportamiento de la curva de escala en muestras estandarizadas generadas a partir de distintas distribuciones junto con la curva esperada para la distribución normal bivariante. Estos ejemplos se muestran a través de la Figuras 4.1 a 4.4, en las que se representan, para tamaños muestrales 50 y 100, curvas muestrales y esperadas obtenidas con la profundidad semiespacial. La curva en color rojo es la curva de escala esperada para la distribución normal estándar bivariante y el tamaño muestral correspondiente. Las curvas en color azul representan las curvas de escala de diez muestras generadas a partir de distintas distribuciones. La distribución generadora para la Figura 4.1 es la normal estándar. Las distribuciones de las otras tres figuras consiste en un vector bivariante con sus dos coordenadas independientes e igualmente distribuidas. En la Figura 4.2 cada

coordenada del vector ha sido generada a partir de una distribución exponencial, en la 4.3 a partir de una distribución t con dos grados de libertad y en la 4.4 a partir de una uniforme.

En primer lugar, se puede observar cómo la variabilidad en las curvas de escala muestrales se reduce al aumentar el tamaño muestral. En segundo lugar, puede verse cómo para las muestras no normales, las diferencias entre curvas muestrales y esperadas (bajo normalidad) son menores en el caso exponencial (Figura 4.2), en el que se observa además cómo para regiones centrales pequeñas el volumen muestral es menor que para la distribución normal. Esto es así debido a que posee una mayor curtosis, es decir, como tiene muchos puntos concentrados en poco espacio es necesaria una cantidad importante de éstos para que el volumen empiece a aumentar. Esto puede verse más claramente en la Figura 4.2(b) en la que sistemáticamente (aproximadamente hasta $p = 0.7$) todas las curvas muestrales están por debajo de lo que se esperaría para una normal.

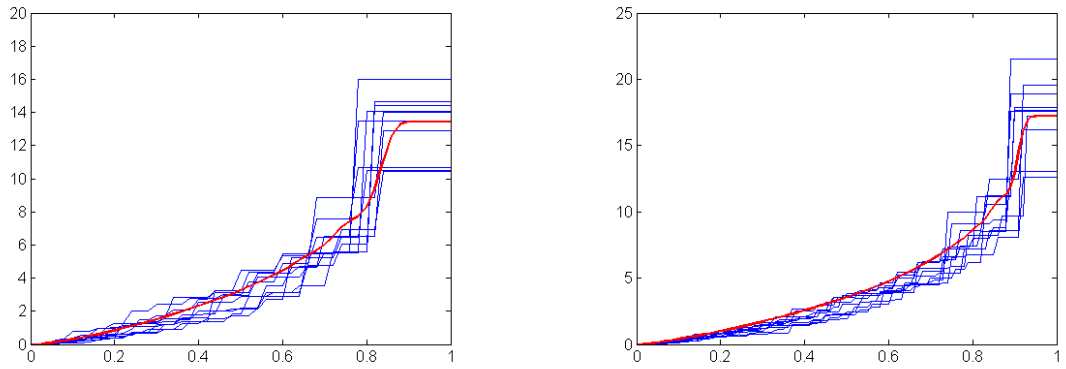


(a) $n = 50$ (b) $n = 100$

Figura 4.1: *Ejemplos de curvas de escala de muestras normales y esperada bajo normalidad para la profundidad semiespacial.*

En cuanto a las otras dos distribuciones, Figuras 4.3 y 4.4, se observa que las discrepancias entre curva muestral y curva esperada son mayores que para el caso exponencial. Se tiene que, para la distribución t de dos grados de libertad, las curvas muestrales están para la mayoría de valores de p por debajo de la esperada y para el resto por encima de dicha curva, presentando incluso valores sustancialmente mayores para la envolvente

convexa de todos los puntos ($p = 1$). Esto se debe a que es una distribución de colas más pesadas que la normal, lo que hace que presente una mayor cantidad de puntos muy alejados de la media y tenga un volumen de envolvente convexa de todos los puntos mucho más elevado. En el caso de coordenadas uniformes se tiene un crecimiento constante que, si bien nunca está demasiado alejado de la curva esperada de la normal debido a que no puede tener valores extremadamente grandes, la mayor parte del tiempo está separada de ésta.

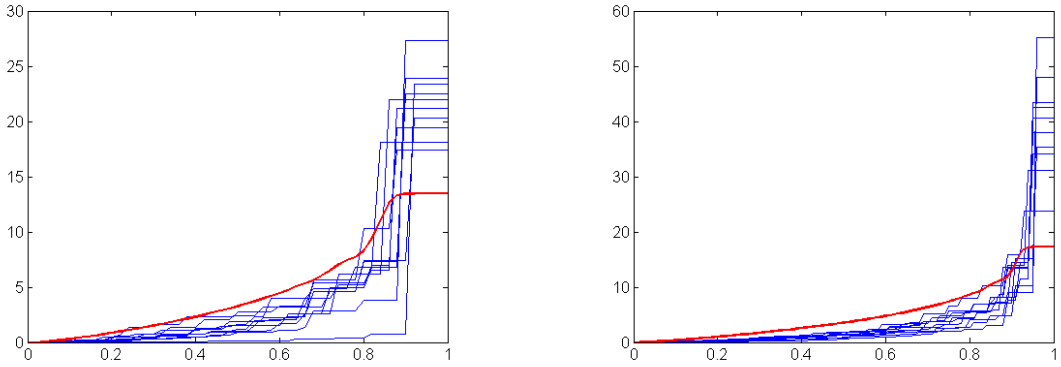


(a) $n = 50$ (b) $n = 100$

Figura 4.2: *Ejemplos de curvas de escala muestrales para diez muestras de dos exponenciales independientes y curva de volumen bajo normalidad. Ordenación según la profundidad semiespacial.*

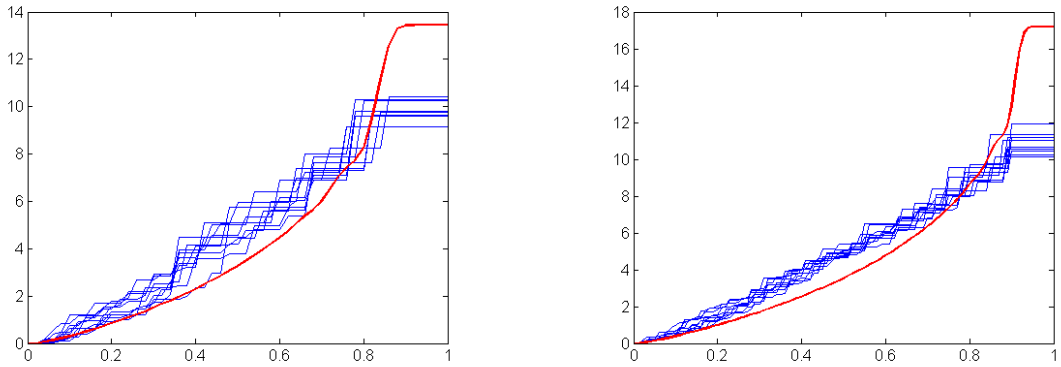
4.2.1. Envolventes convexas

A continuación se analizan las características de las envolventes convexas que se utilizan. En todas las curvas de escala esperadas de las figuras anteriores se observa una particularidad al final de la misma: es constante para un intervalo de valores de p altos. Este hecho se debe a la profundidad escogida para la ordenación, la semiespacial, en la que todos los puntos de la envolvente convexa toman el mismo valor de profundidad. La mayor o menor longitud de este intervalo representa el porcentaje de puntos sobre el total, que forma parte de la envolvente convexa de toda la muestra. Cuanto mayor es el tamaño muestral menor es dicho porcentaje. Esto puede verse tanto en estas figuras, al



(a) $n = 50$ (b) $n = 100$

Figura 4.3: *Ejemplos de curvas de escala muestrales para diez muestras de dos t_2 independientes y curva de volumen bajo normalidad. Ordenación según la profundidad semiespacial.*



(a) $n = 50$ (b) $n = 100$

Figura 4.4: *Ejemplos de curvas de escala muestrales para diez muestras de dos distribuciones uniformes independientes y curva de volumen bajo normalidad. Ordenación según la profundidad semiespacial.*

pasar de 50 a 100 observaciones, como en la Figura 4.5 que contiene las curvas esperadas para la normal bivariante estándar para tamaños muestrales 50, 100, 200, 500 y 1000, en la cual el intervalo constante es casi inapreciable para la curva de tamaño muestral 1000 (curva en color verde).

En Hueter (1999) puede encontrarse la siguiente cota superior para la esperanza del

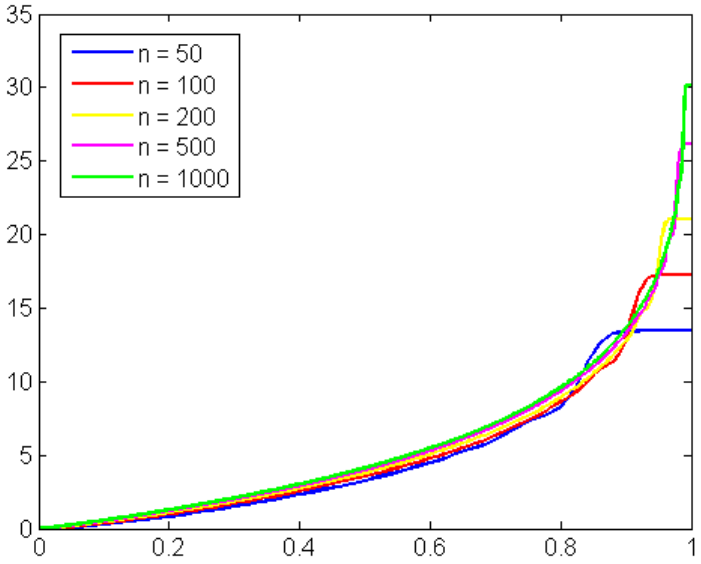


Figura 4.5: *Curvas de escala esperadas para muestras de diferentes tamaños y distribución normal bivariante. Ordenación según la profundidad semiespacial.*

número de vértices de la envovente convexa para muestras de n observaciones de distribuciones normales d-dimensionales (N_n),

$$E\left[N_n\right] \leq c\left(ln\left(n\right)\right)^{(d-1)/2},$$

donde $c = 2\sqrt{d-1}\left(2\pi\right)^{(d-1)/2}/\Gamma\left(d/2\right)$ y $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$. Gracias a esta cota se tiene que la esperanza del porcentaje de puntos en la envolvente convexa para muestras normales tiende a cero cuando el tamaño muestral aumenta,

$$\lim_{n\to\infty}\frac{E\left[N_n\right]}{n} \leq lim_{n\to\infty}\frac{c\left(ln\left(n\right)\right)^{(d-1)/2}}{n} = 0.$$

Otra característica de la curva de escala es que depende de la función de profundidad escogida para realizar la ordenación. En la Figura 4.6 se representa la curva de escala esperada para muestras de tamaño 100, simuladas a partir de una normal bivariante estándar y con ordenaciones obtenidas a través de distintas profundidades. Se puede observar que las profundidades de Oja, L_1 y por proyecciones poseen curvas prácticamente iguales. Se puede comprobar también que en la parte final de la curva las diferencias

entre profundidades son más significativas, ya que las curvas de estas tres profundidades junto con la de bandas modificada presentan mayor suavidad y no tienen una parte final constante. El hecho de poseer una parte final constante depende de la variedad de posibles valores de la profundidad sobre una muestra. En la profundidad simplicial ocurre lo mismo que en la semiespacial, los puntos de la envolvente convexa poseen el mismo valor de profundidad y forman, por lo tanto, una clase de equivalencia que determina la forma del final de la curva. Para la profundidad por bandas se producen también numerosos empates para valores de profundidad bajos, aunque su cantidad es menor que para estas dos profundidades.

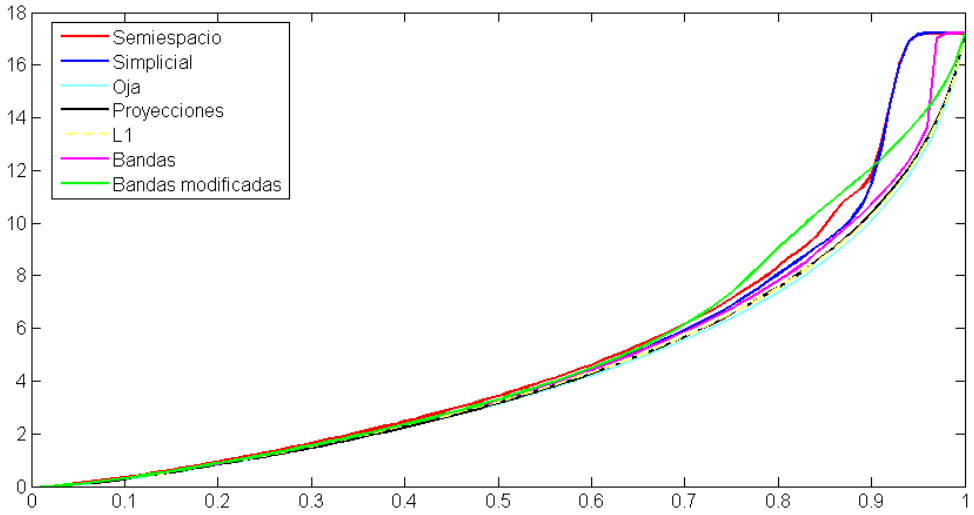


Figura 4.6: *Curva media para muestras de tamaño 100 de una normal estándar bivariante para varias profundidades.*

Como ya se ha comentado, el patrón de crecimiento de la curva de escala para normales con matriz de covarianzas diferentes es equivalente. Este hecho puede observarse en la Figura 4.7, en la que se representan las curvas medias para muestras de tamaño 100 para cinco distribuciones normales con matrices de varianzas covarianzas diferentes. Las ordenaciones para este ejemplo han sido obtenidas según la profundidad semiespacial. Todas las curvas de la figura son proporcionales y el factor de proporcionalidad con respecto a la curva de la normal estándar es igual a la desviación típica generalizada de

cada distribución, la cual dada una variable d-dimensional X con matriz de covarianzas Σ, se define como $(\det(\Sigma))^{1/2}$. Tomando como referencia la curva de la normal estándar se tiene que la curva roja es el doble y la verde, magenta y amarilla son, respectivamente, $\sqrt{2}$, 0.6 y $\sqrt{3}$ veces la estándar.

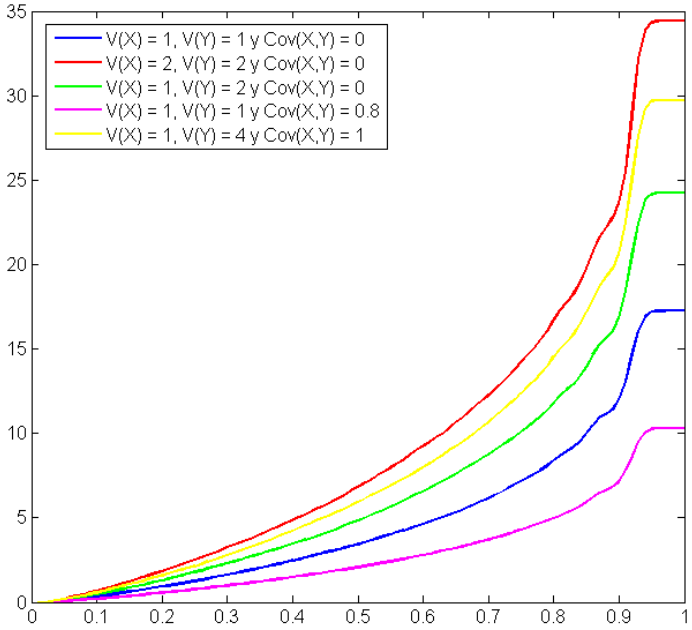


Figura 4.7: *Curvas medias de muestras de tamaño 100 para normales bivariantes de varianzas diferentes según la profundidad semiespacial.*

Proposición 4.1 *Dado el conjunto de puntos $x_1, x_2, \ldots, x_n$ en $\mathbb{R}^d$ $(n > d)$ y dada una matriz de varianzas-covarianzas Σ de dimensión d, se tiene que*

$$Vol\left(env\ conv\left(\Sigma^{1/2}x_1, \Sigma^{1/2}x_2, \ldots, \Sigma^{1/2}x_n\right)\right) = \det(\Sigma)^{1/2}\ Vol(env\ conv(x_1, x_2, \ldots, x_n)).$$

Demostración. Como la envolvente convexa de los n puntos puede ser descompuesta en un conjunto finito y disjunto de envolventes convexas de subconjuntos de $d+1$ elementos, basta con probar que se verifica para esas envolventes de subconjuntos. Sin pérdida de generalidad se toman los primeros $d+1$ puntos del conjunto ya multiplicado

por $\Sigma^{1/2}$. El volumen de la envolvente convexa de estos puntos es igual a

$$\frac{1}{(d+1)!}\left|\det\begin{pmatrix} \Sigma^{1/2}x_1 & \Sigma^{1/2}x_2 & \dots & \Sigma^{1/2}x_{d+1} \\ 1 & 1 & \dots & 1 \end{pmatrix}\right|$$

y, restando la primera columna a todas las demás,

$$\begin{aligned}
&= \frac{1}{(d+1)!}\left|\det\begin{pmatrix} \Sigma^{1/2}x_1 & \Sigma^{1/2}(x_2-x_1) & \dots & \Sigma^{1/2}(x_{d+1}-x_1) \\ 1 & 0 & \dots & 0 \end{pmatrix}\right| \\
&= \frac{1}{(d+1)!}\left|\det\begin{pmatrix} \Sigma^{1/2}(x_2-x_1) & \dots & \Sigma^{1/2}(x_{d+1}-x_1) \end{pmatrix}\right| \\
&= \frac{\det\left(\Sigma^{1/2}\right)}{(d+1)!}\left|\det\begin{pmatrix} x_2-x_1 & \dots & x_{d+1}-x_1 \end{pmatrix}\right| \\
&= \frac{\det(\Sigma)^{1/2}}{(d+1)!}\left|\det\begin{pmatrix} x_1 & x_2-x_1 & \dots & x_{d+1}-x_1 \\ 1 & 0 & \dots & 0 \end{pmatrix}\right| \\
&= \frac{\det(\Sigma)^{1/2}}{(d+1)!}\left|\det\begin{pmatrix} x_1 & x_2 & \dots & x_{d+1} \\ 1 & 1 & \dots & 1 \end{pmatrix}\right| \\
&= \det(\Sigma)^{1/2}\,\mathrm{Vol}\left(\text{env conv}\left(x_1, x_2, \dots, x_{d+1}\right)\right). \blacksquare
\end{aligned}$$

4.2.2. Estadístico del contraste

Este contraste de hipótesis trata de determinar si la curva de escala de un conjunto de datos es igual a la que se obtendría para una determinada función de distribución F_0, es decir,

$$\begin{aligned}
H_0 &: C_{n,p} \text{ es igual a } C_{F_0,p} \\
H_1 &: C_{n,p} \text{ no es igual a } C_{F_0,p}.
\end{aligned}$$

El estadístico del contraste debe considerar las discrepancias entre la curva de volumen muestral ($C_{n,p}$)y la nula ($C_{F_0,p}$). Separaciones elevadas entre estas curvas sugieren que la distribución nula no es adecuada para el conjunto de datos, por lo que cualquier función que mida la distancia entre ambas curvas puede usarse como estadístico del contraste.

La función que se propone es el área entre ambas curvas, es decir,

$$A\left(C_{n,p}\right) = \int_0^1 \left|C_{n,p} - C_{F_0,p}\right| \mathrm{d}p.$$

Aunque es posible emplear cualquier función basada en la norma L_k de la diferencia entre las funciones,

$$\int_0^1 |C_{n,p} - C_{F_0,p}|^k \, \mathrm{d}p,$$

donde $1 \leq k < \infty$, se ha tomado el valor $k = 1$ ya que se obtiene una mayor robustez ante observaciones extremas que produzcan que el volumen tome valores elevados.

Como se puede observar en la Figura 4.5, la diferencia entre la curva muestral para $n = 50$ y la curva para $n = 1000$ para valores de p mayores que 0.8 es elevada. El hecho de que el final de las curvas sea constante para determinadas profundidades hace que no sea adecuado emplear directamente la curva bajo la nula en el estadístico, ya que nunca podría encontrarse una curva muestral con valor del estadístico igual a cero. Para corregir este problema en muestras pequeñas se propone sustituir $C_{F_0,p}$ por una estimación que sí dependa del número de observaciones en la muestra. La estimación que se considera más adecuada es

$$\hat{C}_{F_0,p} = B^{-1} \sum_{i=1}^{B} C_{n,p}\left(X_F^{i)}\right),$$

donde $C_{n,p}\left(X_F^i\right)$ es la curva de escala muestral para la i-ésima muestra $X_F^{i)}$, $X_F^{i)}$ es una muestra aleatoria de tamaño n de F y B es un número suficientemente grande para que la variabilidad de las curva media sea despreciable. Con esta estimación el estadístico del contraste, queda como

$$A\left(C_{n,p}\right) = \int_0^1 \left|C_{n,p} - \hat{C}_{F_0,p}\right| \mathrm{d}p.$$

El estadístico así definido no verifica de forma estricta la propiedad de invarianza ante transformaciones afines bajo la hipótesis nula que cualquier test de bondad de ajuste debería cumplir. A continuación se propone una modificación con la que se consigue la invarianza en los casos en que las ordenaciones se realicen con funciones de profundidad que verifiquen la propiedad de invarianza afín.

La modificación consiste en dividir el estadístico por la desviación típica generalizada de la distribución normal de la hipótesis nula,

$$A'\left(C_{n,p}\right) = \frac{1}{\det\left(\Sigma\right)^{1/2}} \int_0^1 \left|C_{n,p} - \hat{C}_{F_0,p}\right| \mathrm{d}p,$$

donde Σ denota la matriz de varianzas-covarianzas bajo la hipótesis nula. Esta modificación es válida cuando la hipótesis nula es simple. Cuando es compuesta (no se fija el valor de Σ) hay que sustituir la matriz Σ por una estimación suya, por ejemplo, con la cuasivarianza muestral $S = n^{-1}\left(X - \overline{X}\right)'\left(X - \overline{X}\right)$.

Observación 4.1 *Se puede probar de manera inmediata que el resultado de la modificación es equivalente a estandarizar la muestra por la matriz de covarianzas de la hipótesis nula o muestral, según sea dicha hipótesis nula o compuesta, y posteriormente calcular el estadístico* $A\left(C_{n,p}\right)$.

4.2.3. Valores críticos

Debido a la dificultad teórica introducida con las ordenaciones obtenidas con las funciones de profundidad, no ha sido posible encontrar una forma cerrada ni para la curva esperada bajo la nula ni para la distribución muestral del estadístico del contraste, por lo que tanto la curva esperada como los percentiles de dicha distribución han sido estimados mediante simulación.

La Tabla 4.1 contiene una estimación de los percentiles 0.9, 0.95 y 0.99 para el contraste de hipótesis con distribución nula normal, para las funciones de profundidad más relevantes y tamaños muestrales 50 y 100. Éstos se han obtenido mediante simulación, a partir de 5000 muestras estandarizadas de normales con matriz de varianzas identidad. La curva esperada que se emplea en el estadístico $A'\left(C_{n,p}\right)$ se ha estimado también mediante la simulación de 5000 muestras estandarizadas.

Los percentiles estimados para el vector bidimensional con coordenadas uniformes independientes y para el vector bidimensional con coordenadas exponenciales se encuentran en las Tablas 4.2.

4.2.4. Potencia del contraste

El estudio de la potencia para los contrastes de bondad de ajuste presenta una mayor complejidad que, por ejemplo, para los contrastes sobre la media o sobre cualquier otra

		Percentil		
Profundidad	Tamaño muestral	0.90	0.95	0.99
Semiespacio	50	0.90	1.02	1.28
	100	0.67	0.76	0.94
Simplicial	50	0.83	0.96	1.24
	100	0.64	0.72	0.90
Oja	50	0.56	0.64	0.81
	100	0.48	0.54	0.67
Proyecciones	50	0.60	0.68	0.85
	100	0.50	0.57	0.71
L_1	50	0.58	0.65	0.82
	100	0.49	0.56	0.70
Bandas	50	0.65	0.74	0.94
	100	0.53	0.59	0.75
Bandas modificacada	50	0.71	0.81	1.07
	100	0.68	0.78	1.05

Tabla 4.1: *Valores críticos para el contraste de dispersión con distribución nula normal.*

característica de una distribución, ya que en éstos la hipótesis alternativa resulta ser sencilla en comparación con los de bondad de ajuste, en los que la alternativa es cualquier función de distribución. Por lo tanto, para estudiar la potencia en este tipo de contrastes lo máximo que se puede hacer es un análisis sobre un número elevado de distribuciones de formas diferentes para poder asegurar de algún modo que el contraste ha sido probado para alternativas heterogéneas.

A continuación se introducen los resultados de potencia obtenidos mediante simulación para la distribución normal multivariante y los vectores bidimensionales uniformes y exponenciales. Debido a su importacia en Estadística, se hace un mayor énfasis en los resultados para la distribución normal.

4.2.4.1. Potencia para distribución nula normal

En este apartado se presentan los resultados de simulación de la potencia del contraste frente a 36 distribuciones alternativas que presentan una heterogeneidad considerable. Es-

		Uniforme			Exponencial		
		Percentil			Percentil		
Profundidad	Tamaño muestral	0.90	0.95	0.99	0.90	0.95	0.99
Semiespacio	50	0.065	0.075	0.095	1.501	1.865	2.8
	100	0.048	0.056	0.071	1.225	1.451	2.095
Simplicial	50	0.064	0.074	0.092	1.43	1.788	2.563
	100	0.047	0.055	0.072	1.194	1.423	1.973
Oja	50	0.059	0.068	0.089	0.582	0.675	0.931
	100	0.046	0.053	0.070	0.487	0.559	0.748
Proyecciones	50	0.062	0.072	0.094	0.629	0.731	0.969
	100	0.046	0.054	0.070	0.522	0.601	0.756
L_1	50	0.062	0.072	0.093	0.657	0.759	1.025
	100	0.046	0.054	0.068	0.546	0.614	0.78
Bandas	50	0.062	0.072	0.093	0.982	1.18	1.731
	100	0.047	0.054	0.068	0.828	0.961	1.289
Bandas modificada	50	0.061	0.072	0.094	1.203	1.515	2.469
	100	0.046	0.053	0.069	1.281	1.616	2.575

Tabla 4.2: *Valores críticos para el contaste de dispersión con distribución nula uniforme y exponencial.*

tos ejemplos de distribuciones alternativas se encuentran en el estudio de simulación del contraste de bondad de ajuste basado en distancias entre puntos propuesto en Bartoszynski et al. (1997). Estas 36 distribuciones alternativas se han separado en cuatro grupos. El primero está compuesto por distribuciones bivariantes en las que cada coordenada del vector bidimensional se distribuye según una distribución univariante y donde ambas coordenadas son independientes. En el segundo grupo se encuentran mixturas de normales con diferentes medias y matrices de covarianzas. El tercer grupo está compuesto por distribuciones de Pearson y esféricamente simétricas y el cuarto por distribuciones con correlación radial/angular.

Las tablas presentan el porcentaje de rechazos de la hipótesis nula de normalidad para un contraste de nivel de significación $\alpha = 0.95$ y para las profundidades cuyos valores críticos aparecen en la tabla 4.1. El porcentaje se ha obtenido mediante simulación de 1000 muestras estandarizadas de tamaños muestrales 50 y 100 de la distribución alternativa.

Para realizar una comparación de la potencia del contraste para las diferentes profundidades empleadas en la ordenación de las muestras se obtiene un índice. Para cada una de las distribuciones alternativas se ordenan los valores de potencia y se asigna el valor uno a la profundidad que tiene un valor más alto y el siete a la que tiene el menor valor. Si dos profundidades tienen el mismo valor se les asigna el valor igual al número de profundidades menor que éstas más uno. El índice para cada profundidad se obtiene promediando las puntuaciones sobre todas las distribuciones alternativas. Se calcula un índice por cada uno de los cuatro grupos de alternativas y otro global, para analizar sobre qué tipos de distribuciones se comportan mejor unas profundidades que otras.

La Tabla 4.3 contiene los resultados de potencia para las distribuciones compuestas por dos distribuciones univariantes iguales e independientes. Se puede observar cómo aumenta la potencia cuando la alternativa presenta una mayor asimetría o cuando presenta unas colas más pesadas. También puede notarse que para algunas distribuciones las diferencias entre profundidades son elevadas, como en el caso de la chi-cuadrado con cinco grados de libertad y la exponencial. Esto se debe a que en distribuciones más asimétricas se tiene una mayor diferencia al ordenar la muestra. Los casos en los que todas las profundidades se comportan de forma homogénea y con un porcentaje alto de rechazo son la beta(1,1) (es decir, $U(0,1)$), la lognormal y la t de Student de dos grados de libertad.

Cuando una de las coordenadas del vector bivariante es la normal estándar, Tabla 4.4, la potencia se reduce en todos los casos, pero se mantenienen como mejores los casos de la exponencial y la beta(1,1), para los que la potencia sigue siendo mayor del cincuenta por ciento para la mayoría de las profundidades.

En la Tabla 4.5 aparecen el índice calculado para todas las distribuciones de las Tablas 4.3 y 4.4. En gris oscuro aparece, para cada tamaño muestral, el valor de la función de profundidad que mejor se ha comportado de forma global y en gris claro la segunda mejor. Aunque intercambian posiciones conforme el tamaño muestral aumenta, las dos profundidades que más potencia ofrecen son la de Oja y la L_1.

El segundo grupo de distribuciones contiene mixturas al cincuenta por ciento de dos distribuciones normales bivariantes. Cada ejemplo de mixtura se denota por la tupla

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
Exponencial	50	25	24	99	99	91	58	48
	100	51	50	100	100	100	91	72
Lognormal	50	73	75	100	100	100	95	90
	100	98	97	100	100	100	100	99
Gamma(5,1)	50	8	8	18	18	12	11	11
	100	12	13	36	35	21	15	16
chi-cuadrado(5)	50	9	11	52	50	33	16	17
	100	18	20	89	89	66	33	26
chi-cuadrado(15)	50	7	7	12	11	11	7	9
	100	10	10	17	20	15	10	11
t(2)	50	97	97	97	97	98	97	96
	100	100	100	100	100	100	100	100
t(5)	50	52	52	45	45	45	50	51
	100	78	77	76	74	74	76	75
Logística(0,1)	50	27	28	27	24	23	25	29
	100	50	47	43	42	44	46	46
Beta(1,1)	50	87	90	86	79	89	90	78
	100	100	100	100	100	100	100	100
Beta(1,2)	50	40	45	28	25	40	40	28
	100	90	91	86	84	82	86	74
Beta(2,2)	50	26	27	31	31	39	35	20
	100	75	76	78	73	76	77	60

Tabla 4.3: *Potencia del contraste de dispersión con distribución nula normal para el grupo de alternativas de coordenadas independientes e idénticamente distribuidas.*

(a, b, c), que representa la mixtura de las distribuciones

$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & b \\ b & 1 \end{pmatrix}\right) \; y \; N\left(\begin{pmatrix} a \\ a \end{pmatrix}, \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}\right).$$

La potencia simulada para las mixturas se presenta en la tabla 4.6. Puede observarse que el contraste para una muestra de 50 observaciones detecta en torno a un diez por ciento si los componentes del vector de medias se separan en dos unidades, incrementándose ligeramente cuando se modifica la matriz de covarianzas de una de ellas. La

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
Normal(0,1) y Exponencial	50	11	14	67	64	48	27	23
	100	23	23	97	97	81	50	34
Normal(0,1) y chi-cuadrado(5)	50	8	8	18	21	15	11	12
	100	11	13	35	39	22	15	13
Normal(0,1) y t(5)	50	23	23	20	22	19	24	23
	100	43	39	34	34	35	37	37
Normal(0,1)y Beta(1,1)	50	20	23	23	21	32	28	16
	100	60	66	64	57	67	64	43
Normal(0,1) y Beta(1,2)	50	9	10	9	7	12	10	6
	100	26	27	21	16	25	26	16

Tabla 4.4: *Potencia del contraste de dispersión con distribución nula normal, para el grupo de alternativas de coordenadas independientes no idénticamente distribuidas.*

	Profundidad de ordenación						
n	PSem	PS	PO	PP	PL_1	PB	PBM
50	5.09	4.25	3.22	3.81	3.09	3.59	4.94
100	4.44	4.09	3.09	3.94	3.66	3.72	5.06

Tabla 4.5: *Índice de rangos para el contraste de dispersión con distribución nula normal y el grupo 1 de alternativas.*

potencia está en torno a un cuarenta por ciento si la diferencia es de cuatro unidades y sube al cincuenta por ciento cuando ambas normales tienen correlación elevada y de signo contrario.

De nuevo se observa (véase Tabla 4.7) que, de forma global para el grupo, las profundidades que mejor se comportan son la de Oja y L_1 y, para tamaño muestral 100, también la profundidad por bandas.

El tercer grupo de distribuciones está formado por distribuciones esféricamente simétricas, de Pearson tipos II y VII y esféricas cuyo radio sigue una distribución determinada. La Tabla 4.8 contiene las estimaciones de la potencia frente a estas alternativas. Se observa que a diferencia de los grupos anteriores, todas las profundidades se comportan de

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
Mixtura Normal (2,0,0)	50	7	9	12	9	13	10	6
	100	17	21	24	20	25	28	13
Mixtura Normal (4,0,0)	50	34	42	38	37	54	47	36
	100	84	92	87	86	93	95	82
Mixtura Normal (2,0.9,0)	50	9	8	24	27	15	13	10
	100	31	24	56	52	34	36	15
Mixtura Normal (0.5,0.9,0)	50	24	24	28	33	22	24	22
	100	41	42	45	49	45	42	24
Mixtura Normal (0.5,0.9,-0.9)	50	42	44	57	53	55	51	62
	100	85	77	91	83	90	89	95

Tabla 4.6: *Potencia del contraste de dispersión con distribución nula normal, para el grupo de alternativas formado por mixturas de normales.*

	Profundidad de ordenación						
n	PSem	PS	PO	PP	PL_1	PB	PBM
50	6.00	4.90	2.40	3.10	2.90	3.60	5.10
100	5.60	4.90	2.50	3.80	2.70	2.70	5.80

Tabla 4.7: *Índice de rangos para el contraste de dispersión con distribución nula normal y el grupo 2 de alternativas.*

forma muy homogénea sobre la misma distribución. La potencia para las distribuciones de Pearson es mayor que para las distribuciones esféricas de radio aleatorio, sobre las que sobresalen las que tienen por radio la distribución exponencial y la beta(1,1).

En la Tabla 4.9 se presentan los índices para este grupo. Dependiendo del tamaño muestral se muestran mejores unas que otras. Si éste es bajo destaca en primer lugar la profundidad L_1 y en segundo la de Oja. Cuando aumenta, la profundidad por bandas es la mejor seguida de la de Oja.

El último grupo de alternativas está compuesto por la distribución definida en Quiroz y Dudley (1991), la cual presenta una correlación entre las componentes radial y angular de las coordenadas estandarizadas. Se emplean diferentes valores para la correlación. En

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
PearsonII(0)	50	94	96	91	85	97	96	84
	100	100	100	100	100	100	100	100
PearsonII(1)	50	41	43	49	47	52	49	26
	100	92	94	92	89	93	92	70
PearsonVII(2)	50	99	98	99	99	98	98	97
	100	100	100	100	100	100	100	100
PearsonVII(3)	50	75	77	72	72	73	74	70
	100	95	95	95	94	95	96	89
PearsonVII(5)	50	34	35	33	33	31	34	29
	100	62	59	55	55	55	59	48
Esférica(Exponencial)	50	100	100	100	100	100	99	100
	100	100	100	100	100	100	100	100
Esférica(Gamma(5,1))	50	7	7	12	10	15	13	8
	100	18	19	28	25	28	28	18
Esférica(Beta(1,1))	50	11	14	38	39	40	30	16
	100	75	79	88	84	85	85	21
Esférica(Beta(1,2))	50	60	56	74	73	74	65	57
	100	88	87	95	94	94	94	77
Esférica(Beta(2,2))	50	18	21	27	26	29	26	11
	100	64	70	66	60	64	66	34

Tabla 4.8: *Potencia del contraste de dispersión con distribución nula normal, para el grupo de alternativas formado por distribuciones de Pearson y esféricas.*

	Profundidad de ordenación						
n	PSem	PS	PO	PP	PL_1	PB	PBM
50	4.45	4.25	3.25	3.80	2.50	3.60	6.15
100	4.25	3.60	3.10	4.60	3.45	2.95	6.05

Tabla 4.9: *Índice de rangos para el contraste de dispersión con distribución nula normal y el grupo 3 de alternativas.*

la Tabla 4.10 se puede observar cómo a mayor correlación, mayor es el porcentaje de rechazos. Como se ve en la Tabla 4.11, las profundidades que mejor se comportan ante esta alternativa, para muestras pequeñas son la de proyecciones, Oja y L_1 y para muestras

grandes la de Oja y proyecciones.

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
0.2	50	5	4	5	5	6	6	6
	100	4	6	6	5	5	7	5
0.4	50	4	5	8	9	7	5	6
	100	7	8	10	10	7	9	5
0.6	50	7	8	10	19	10	9	8
	100	15	14	19	28	14	16	9
0.8	50	17	17	36	44	19	17	11
	100	42	36	69	80	40	37	19
1	50	41	40	92	95	57	47	35
	100	93	93	100	100	99	96	86

Tabla 4.10: *Potencia del contraste de dispersión con distribución nula normal, para el grupo de alternativas con correlación radial/angular.*

	Profundidad de ordenación						
n	PSem	PS	PO	PP	PL_1	PB	PBM
50	5.80	5.80	2.70	1.80	2.70	4.10	5.10
100	5.00	4.70	1.90	2.00	4.60	3.20	6.60

Tabla 4.11: *Índice de rangos para el contraste de dispersión con distribución nula normal y el grupo 4 de alternativas.*

De forma global para todas las alternativas (véase la Tabla 4.12), las ordenaciones que mejor comportamiento presentan para muestras de tamaño 50 son, por orden decreciente, las basadas en las profundidades L_1, de Oja, por proyecciones y por bandas y, para muestras de tamaño 100, la de Oja, por bandas, L_1 y por proyecciones.

4.2.4.2. Potencia para distribución nula uniforme

En esta sección se presentan los resultados de potencia en el caso en que se quiera contrastar si la curva de escala es igual a la de una distribución uniforme en el cuadrado $[0,1] \times [0,1]$.

	Profundidad de ordenación						
n	PSem	PS	PO	PP	PL_1	PB	PBM
50	5.14	4.56	3.04	3.43	2.85	3.67	5.32
100	4.63	4.15	2.85	3.83	3.60	3.29	5.65

Tabla 4.12: *Índice de rangos global para el contraste de dispersión con distribución nula normal.*

Las distribuciones alternativas para las que se analiza la potencia del contraste se enumeran a continuación. Algunas de las alternativas se corresponden a distribuciones en determinados recintos, mientras que otras son las distribuciones de vectores aleatorios bidimensionales, cuyas coordenadas son independientes e idénticamente distribuidas según una distribución unidimensional.

1. Vector con coordenadas distribuidas según beta con parámetros $\alpha = \beta = 0.8$, $\alpha = \beta = 0.9$, $\alpha = \beta = 1.15$ y $\alpha = \beta = 1.3$.

2. Distribución uniforme en la circunferencia unidad.

3. Distribución normal estándar y normal estándar truncada fuera de las circunferencias de radio 1, 1.5 y 2.

4. Mixtura de uniformes en cuadrados: $\alpha U\left[0,1\right]^2 + (1-\alpha)\, U\left[0,b\right]^2$, con valores $\alpha = 0.9$ y $b = 0.25$, $\alpha = 0.75$ y $b = 0.25$, $\alpha = 0.9$ y $b = 0.5$ y $\alpha = 0.75$ y $b = 0.5$.

5. Distribución uniforme en la región resultante de intersecar el cuadrado $[0,1] \times [0,1]$ con el semiespacio $y \geq b - x$, para valores de b iguales a 0.2, 0.4, 0.6, 0.8 y 1.

6. Vector con coordenadas distribuidas según una Pearson de tipo II con parámetro igual a 0 y a 1.

La potencia simulada, para la distribución beta, la uniforme en la circunferencia y para las normales se encuentra recogida en la Tabla 4.13. Puede observarse cómo las diferencias entre profundidades no son tan elevadas como para la distribución nula normal.

Las profundidades que presentan valores de potencia más elevados son la semiespacial, simplicial y por bandas.

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
Beta(0.8,0.8)	50	26	24	24	20	25	27	28
	100	49	52	47	45	53	54	50
Beta(0.9,0.9)	50	9	10	9	9	10	9	10
	100	14	14	14	14	18	15	15
Beta(1.15,1.15)	50	9	8	10	9	10	10	8
	100	20	18	17	18	19	20	19
Beta(1.3,1.3)	50	25	24	19	17	22	27	20
	100	52	53	47	45	51	50	52
Unif. Circunferencia	50	33	31	25	26	25	30	22
	100	76	76	67	72	71	76	72
Normal	50	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100
Normal circ. (1)	50	58	56	49	47	49	55	47
	100	95	96	92	93	94	95	92
Normal circ. (1.5)	50	83	79	77	73	78	82	77
	100	100	100	99	100	100	100	100
Normal circ. (2)	50	96	95	94	91	95	96	94
	100	100	100	100	100	100	100	100

Tabla 4.13: *Potencia del contraste de dispersión con distribución nula uniforme y alternativas beta, uniforme en circunferencia y normal.*

Los resultados de potencia para el resto de distribuciones alternativas se encuentran en la Tabla 4.14. Para las mixturas de uniformes, los mejores resultados se dan para la profundidad de Oja y por proyecciones. Mientras que para las alternativas de uniformes en cuadrados recortados y de Pearson, las mejores profundidades son la del semiespacio, simplicial y por bandas.

De forma global (Tabla 4.15), sobre las veinte alternativas se observa que, sobre el porcentaje medio de rechazo, cuando la muestra está formada por 50 observaciones las mejores profundidades son la del semiespacio y la de bandas, mientras que para muestras

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
Mixt. Unif. (0.1,0.5)	50	6	6	7	7	6	5	7
	100	5	5	8	10	5	5	7
Mixt. Unif. (0.25,0.5)	50	17	14	25	27	13	13	11
	100	19	18	41	51	23	19	22
Mixt. Unif. (0.1,0.25)	50	7	8	7	9	6	8	8
	100	8	10	8	9	9	9	12
Mixt. Unif. (0.25,0.25)	50	10	9	32	41	9	10	10
	100	10	8	50	65	12	9	13
Cuad. Recort (0.2)	50	6	5	5	6	5	7	5
	100	6	5	6	5	5	5	5
Cuad. Recort (0.4)	50	12	9	10	9	9	9	9
	100	21	20	19	20	22	18	17
Cuad. Recort (0.6)	50	48	41	34	35	35	43	34
	100	85	83	75	77	78	83	81
Cuad. Recort (0.8)	50	96	93	89	88	88	91	89
	100	100	100	100	100	100	100	100
Cuad. Recort (1)	50	100	100	99	99	98	100	99
	100	100	100	100	100	100	100	100
Pearson II (0)	50	32	29	28	26	24	31	20
	100	76	78	70	71	71	77	70
Pearson II (1)	50	88	88	87	80	88	88	84
	100	100	100	100	100	100	100	100

Tabla 4.14: *Potencia del contraste de dispersión con distribución nula uniforme y alternativas mixtura de uniforme, uniforme en cuadrados recortados y Pearson II.*

de tamaño 100 lo son la de proyecciones y de Oja. En cuanto al índice de la posición media sobre todas las alternativas, se tiene para ambos tamaños muestrales que las mejores son, por este orden, la del semiespacio y la de bandas.

4.2.4.3. Potencia para distribución nula exponencial

Para finalizar el estudio de la potencia del contraste, se presentan los resultados del porcentaje de rechazos para distribución nula exponencial y el conjunto de distribuciones

		Profundidad de ordenación						
	n	PSem	PS	PO	PP	PL_1	PB	PBM
Potencia media	50	43.05	41.45	41.5	40.95	39.75	42.05	39.10
	100	56.10	56.10	57.30	59.05	55.65	56.00	55.60
Rango medio	50	2.60	3.65	4.40	4.65	4.72	3.05	4.93
	100	3.63	3.75	4.88	4.20	3.80	3.73	4.03

Tabla 4.15: *Porcentaje medio de rechazo e índice de rangos para el contraste de dispersión sobre distribución nula uniforme.*

alternativas, que son vectores aleatorios con coordenadas independientes distribuidas según:

1. Distribución normal estándar.
2. $|X|$, donde X tiene distribución normal estándar.
3. Distribución lognormal estándar.
4. Distribución chi-cuadrado con 1, 3, 4, 5 y 10 grados de libertad.
5. Distribución gamma de parámetros $(5, 1)$.
6. Distribución Weibull de parámetros $(1, 0.5)$, $(1, 0.75)$, $(1, 0.9)$, $(1, 1.1)$, $(1, 1.3)$ y $(1, 1.7)$.
7. Mixtura de exponenciales (α,$1/\lambda$): $\alpha exp\,(1) \times exp\,(1) + (1-\alpha)\, exp\,(1/\lambda) \times exp\,(1/\lambda)$, con valores de α iguales a 0.7, 0.8 y 0.9 y valores de λ iguales a 2 y 5.

Los resultados de la potencia para las cinco primeras se presentan en la Tablas 4.16. Se observa, como era esperable, que todas las profundidades rechazan el 100 % de las veces la nula exponencial para muestras normales. Lo mismo sucede para la chi-cuadrado con 10 grados de libertad y para la distribución gamma de parámetros 5 y 1. Según las diferencias que se encuentran para la chi-cuadrado de 1, 3 y 4 grados de libertad, para la lognormal y para el valor absoluto de la normal, las profundidades L_1, de Oja y por bandas son las más potentes para este grupo de alternativas.

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
Normal	50	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100
|Normal|	50	32	30	42	36	46	48	22
	100	91	90	88	86	91	93	62
Lognormal	50	45	45	58	50	53	52	44
	100	66	63	81	80	80	74	64
Chi-cuadrado (1)	50	89	90	93	94	97	96	82
	100	100	100	100	100	100	100	94
Chi-cuadrado (3)	50	12	14	31	26	41	31	12
	100	59	56	69	61	84	74	28
Chi-cuadrado (4)	50	75	71	94	88	96	92	65
	100	100	99	100	100	100	100	91
Chi-cuadrado (5)	50	98	97	100	100	100	100	96
	100	100	100	100	100	100	100	100
Chi-cuadrado (10)	50	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100
Gamma (5.1)	50	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100

Tabla 4.16: *Potencia del contraste de dispersión con distribución nula exponencial y alternativas normal, lognormal, gamma y chi-cuadrado.*

Para las alternativas Weibull y mixturas de exponenciales (Tabla 4.16), se observa que, cuando el parámetro de forma de la distribución Weibull está alejado de 1 (valores 0.5 y 1.7), para todas las profundidades el porcentaje de rechazos es igual a 100 o muy próximo a este valor. La potencia se reduce cuando está próximo a 1, hasta el punto de que la potencia llega a ser menor que la significación del contraste para valor del parámetro igual a 1.1 (esto se debe a un problema de precisión por el tamaño muestral empleado para la estimación de los porcentajes). Para las mixturas de exponenciales, el porcentaje de rechazo es más del doble cuando el parámetro λ de la distribución de contaminación es 5 que cuando es 2 (exponenciales de medias 0.2 y 0.5). También se tiene que para valores de contaminación pequeños, el hecho de doblar el número de

observaciones no contribuye de forma sustancial al aumento de potencia. Esto podría ser debido a la naturaleza asimétrica de la distribución exponencial.

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
Weibull (1.0.5)	50	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100
Weibull (1.0.75)	50	74	73	78	75	82	81	69
	100	94	95	95	95	97	96	83
Weibull (1.0.9)	50	21	22	21	21	21	22	19
	100	32	30	26	28	33	32	23
Weibull (1.1.1)	50	2	2	5	3	4	3	1
	100	7	6	7	7	12	9	4
Weibull (1.1.3)	50	28	26	39	32	54	49	22
	100	80	80	84	79	93	91	53
Weibull (1.1.7)	50	100	99	100	100	100	100	98
	100	100	100	100	100	100	100	100
Mixt. Expo. (0.7.0.2)	50	46	41	57	58	59	53	40
	100	67	65	83	84	86	80	49
Mixt. Expo. (0.8.0.2)	50	24	23	33	30	36	32	23
	100	38	34	45	52	56	46	27
Mixt. Expo. (0.9.0.2)	50	12	11	12	14	13	12	11
	100	14	14	17	19	20	16	14
Mixt. Expo. (0.7.0.5)	50	13	15	16	13	14	14	11
	100	22	17	20	22	19	19	14
Mixt. Expo. (0.8.0.5)	50	11	10	12	10	12	12	8
	100	12	14	15	12	14	12	11
Mixt. Expo. (0.9.0.5)	50	6	9	6	6	7	7	9
	100	8	7	7	8	8	9	7

Tabla 4.17: *Potencia del contraste de dispersión con distribución nula exponencial y alternativas Weibull y mixtura de exponenciales.*

De forma global sobre las 21 alternativas se tiene que, la profundidad que mayor porcentaje de rechazo obtiene es la L_1 para ambos tamaños muestrales, seguida de la de bandas y la de Oja, en ese orden. Se extrae la misma conclusiónal analizar el índice de las posiciones sobre todas las alternativas.

		Profundidad de ordenación						
	n	PSem	PS	PO	PP	PL_1	PB	PBM
Potencia media	50	51.82	51.33	57.00	55.05	58.81	57.33	49.14
	100	66.19	65.24	68.43	68.24	71.10	69.10	58.29
Rango medio	50	4.93	4.75	2.98	3.89	2.50	3.05	5.91
	100	4.07	4.68	3.55	3.70	2.73	3.25	6.02

Tabla 4.18: *Porcentaje medio de rechazo e índice de rangos para el contraste de dispersión sobre distribución nula exponencial.*

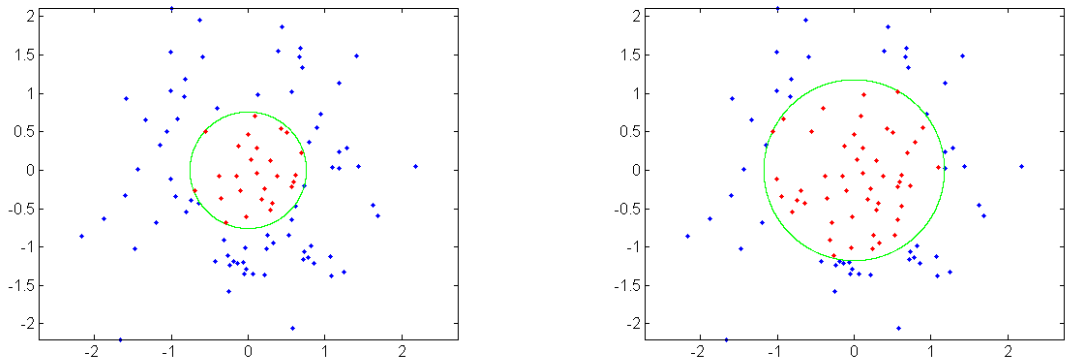
4.3. Contraste basado en las curvas de concordancia

Cuando la dimensión del espacio del conjunto de observaciones es mayor que dos, el tiempo necesario para el cálculo del volumen de la envolvente convexa aumenta de forma exponencial, lo que implica que la aplicación del contraste basado en la curva de escala sea inviable en esta situación. En esta sección se introduce otro contraste basado también en el uso de las regiones centrales que, si bien es computacionalmente costoso, para algunas funciones de profundidad es aplicable en dimensiones mayores que dos.

La idea principal del contraste consiste en la medición, por medio de curvas, de la concordancia de la muestra con la distribución y viceversa. Para medir la concordancia o similitud en cualquiera de las dos direcciones se emplean las regiones centrales de nivel p, tanto muestrales como bajo la hipótesis nula, obteniéndose probabilidades y proporciones de pertenencia a dichas regiones. Con esos valores de pertenencia se construyen las denominadas curvas de concordancia.

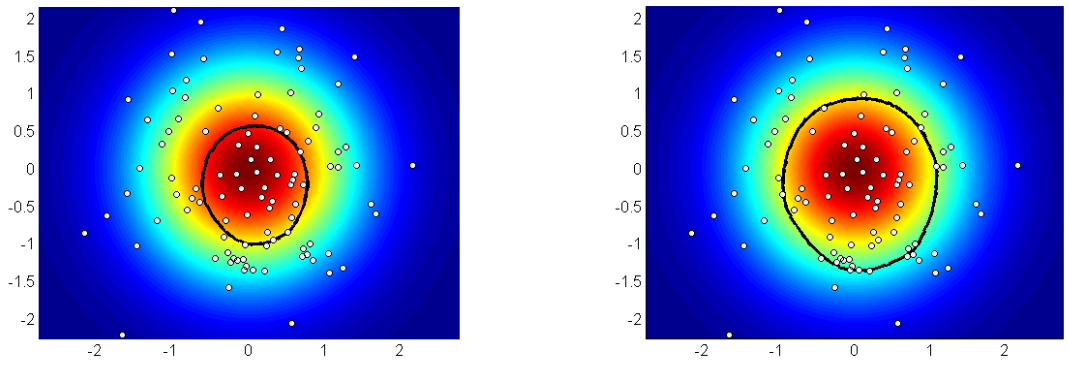
La Figura 4.8 ilustra cómo se estima la concordancia de una muestra con la distribución normal estándar para dos regiones centrales ($p = 0.25$ y $p = 0.50$). El círculo de color rojo representa la región central p-ésima de la normal estándar y los puntos en color rojo representan el conjunto de observaciones de la muestra que están en dicha región. Si la muestra procede de la distribución nula se espera que la proporción de puntos rojos esté en torno a p. La concordancia de la distribución nula con la muestra se ilustra en la Figura 4.9. En ella se representan los contornos de la función de densidad de la normal multivariante, una muestra de tamaño 100 y la región central muestral de niveles 0.25 y

0.50. La concordancia de la nula con la distribución se obtiene calculando la probabilidad bajo la nula de la región central correspondiente. Todas las figuras contenidas en esta sección se han obtenido empleando la profundidad de Oja.



(a) Región central 0.25 (b) Región central 0.50

Figura 4.8: *Regiones centrales 0.25 y 0.50 de una normal estándar y muestra de tamaño 100.*



(a) Región central 0.25 (b) Región central 0.50

Figura 4.9: *Muestra de tamaño 100 de una distribución normal estándar con regiones centrales 0.25 y 0.50 y curvas de nivel de la normal estándar.*

Cuando una función de distribución F es absolutamente continua y no nula en todo el espacio, se tiene que $C_p = R(t_p)$, donde t_p es tal que verifica que $Prob(R(t_p)) = Prob(x \in \mathbb{R}^d : P(x;F) > t_p) = p$. Por lo tanto, para una variable aleatoria X distribuida según F,

se denota por t_p el cuantil $1-p$ de la variable aleatoria $P(X;F)$. Así pues, para obtener la concordancia entre la muestra $x_1, x_2, \ldots, x_n$ y la distribución nula F_0, en vez de obtener C_p y ver cuántos puntos se encuentran en dicha región, se obtiene el cuantil $1-p$ de la variable $P(X;F_0)$ $(P^p(X;F_0))$ y se calcula el porcentaje de puntos de la muestra que tienen, con respecto a la distribución nula, un valor de profundidad mayor que dicho cuantil, es decir,

$$\frac{\#\{x_i : P\left(x_i;F_0\right) > P^{1-p}\left(X;F_0\right)\}}{n}.$$

La obtención de la concordancia en la dirección opuesta se realiza de manera análoga. Dada la muestra $x_1, x_2, \ldots, x_n$, se obtiene el percentil $1-p$ de la muestra $P_n(x_i)$ $(P_n^{1-p}(x))$ y se calcula la probabilidad de que la variable aleatoria $P_n(X)$, con X distribuida según F_0, sea mayor que el percentil, es decir,

$$Prob_{F_0}\left(P_n\left(X\right) > P_n^{1-p}\left(x\right)\right).$$

Definición 4.2 *Sean F_0 la distribución bajo la hipótesis nula y $x_1, x_2, \ldots, x_n$ una muestra aleatoria. Se define la* ***curva de concordancia de muestra con distribución****, como*

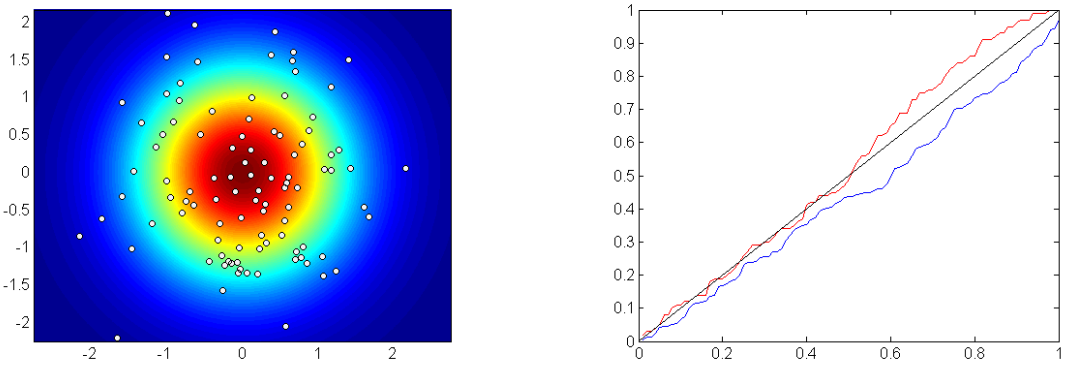
$$CC_n(p) = \frac{\#\{x_i : P\left(x_i;F_0\right) > P^{1-p}\left(X;F_0\right)\}}{n},$$

y de ***distribución con muestra****, como*

$$CC_F(p) = Prob_{F_0}\left(P_n\left(X\right) > P_n^{1-p}\left(x\right)\right).$$

donde $p \in [0,1]$.

A continuación se muestra el comportamiento de estas dos curvas para muestras simuladas de varias distribuciones que se comparan con la distribución normal estándar como nula. Los ejemplos se ilustran en las Figuras 4.10 a 4.16 que contienen, por un lado un diagrama de dispersión de la muestra junto con la función de densidad bajo la nula y, por otro, un gráfico de las curvas de concordancia, donde la curva roja representa la concordancia de muestra con distribución y la azul a concordancia inversa.



(a) Diagrama de dispersión y contornos esperados

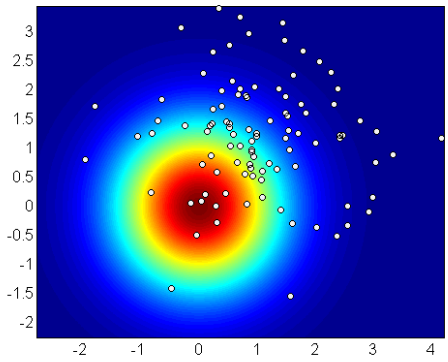
(b) Curvas de concordancia

Figura 4.10: *Comparación entre una muestra de tamaño 100 de una normal estándar con la distribución normal estándar.*

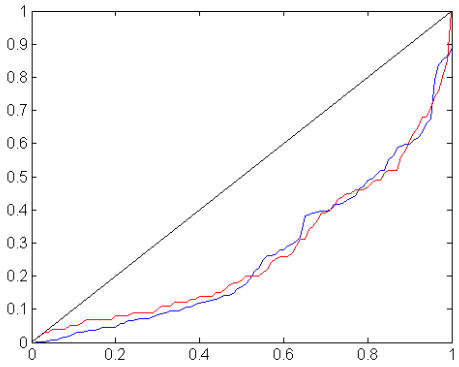
La Figura 4.10 se ha obtenido para una muestra sin estandarizar de la normal estándar. Se observa cómo las dos curvas de concordancia están muy próximas y apenas difieren de la línea en negro que muestra lo que se espera si la muestra tiene distribución F_0.

Las muestras de las figuras 4.11 y 4.12 se han simulado también para una distribución normal, pero con vector de medias $(1, 1)'$ y matriz de covarianzas identidad en el primer caso y con vector de medias cero y matriz de varianza dos veces la identidad en el segundo. Cuando la variación se produce en la media, las dos curvas se alejan de lo esperado bajo la nula y entre ellas apenas hay diferencias. Sistemáticamente las curvas de concordancia están por debajo de lo esperado, ya que los contornos iniciales están separados y apenas tienen intersección. Si la variación es en la variabilidad, la concordancia de la muestra con la distribución se sitúa por debajo de la media, ya que está más dispersa y los contornos de la nula son demasiado pequeños para conterner el número de puntos esperado. La concordancia de la distribución con la muestra presenta un comportamiento opuesto, porque los contornos de la muestra son demasiado grandes en relación a la nula, por lo que la probabilidad de pertenencia es mayor.

Las Figuras 4.13 a 4.16 contienen las curvas en ejemplos no normales. Para una muestra simulada de dos exponenciales independientes y estandarizada (Figura 4.13) se

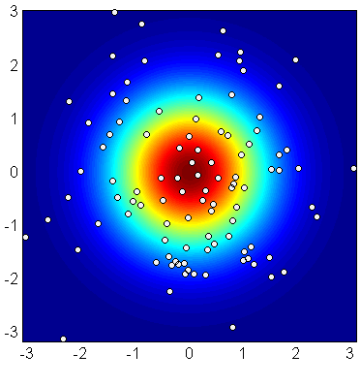


(a) Diagrama de dispersión y contornos esperados

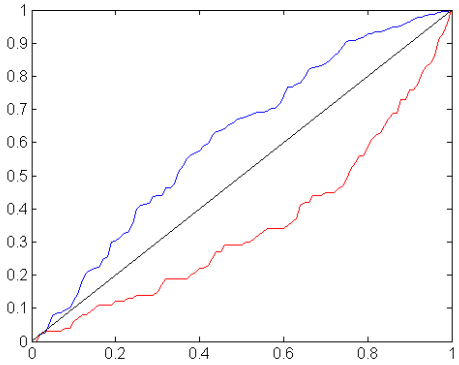


(b) Curvas de concordancia

Figura 4.11: *Comparación entre una muestra de tamaño 100 de una normal con vector de medias (1,1)' y la distribución normal estándar.*



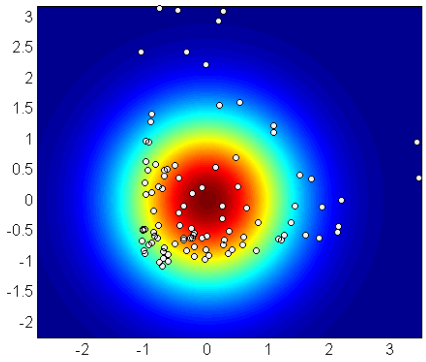
(a) Diagrama de dispersión y contornos esperados



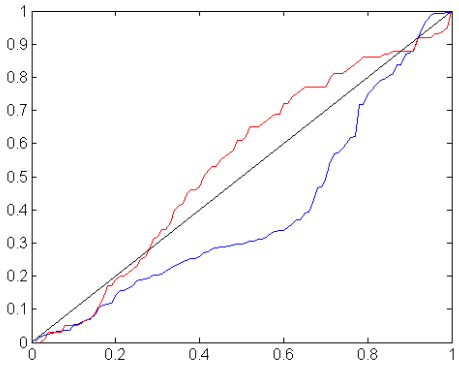
(b) Curvas de concordancia

Figura 4.12: *Comparación entre una muestra de tamaño 100 de una normal con varianza* $\sigma_{11} = \sigma_{22} = 2$ *y* $\sigma_{12} = 0$ *con la distribución normal estándar.*

observa que la concordancia de distribución con muestra detecta de forma más efectiva ese cambio de forma. Si la muestra estandarizada proviene de distribuciones uniformes independientes, Figura (4.14) ninguna de las dos curvas detecta en gran medida la discrepancia entre distribuciones, si bien la distancia con lo esperado es mayor que para la muestra de la normal estándar (véase Figura 4.10).

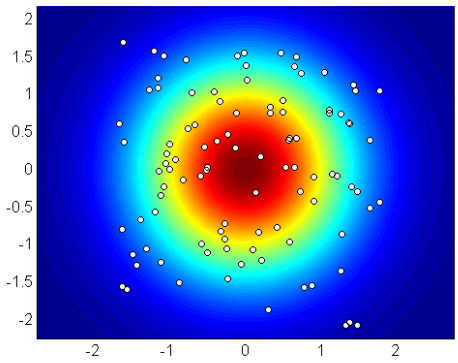


(a) Diagrama de dispersión y contornos esperados

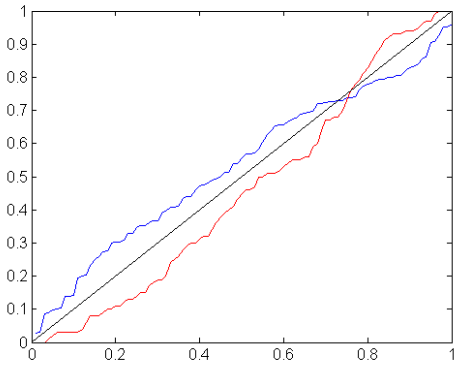


(b) Curvas de concordancia

Figura 4.13: *Comparación entre una muestra estandarizada de tamaño 100 de un vector con componentes exponenciales y la distribución normal estándar.*



(a) Diagrama de dispersión y contornos esperados

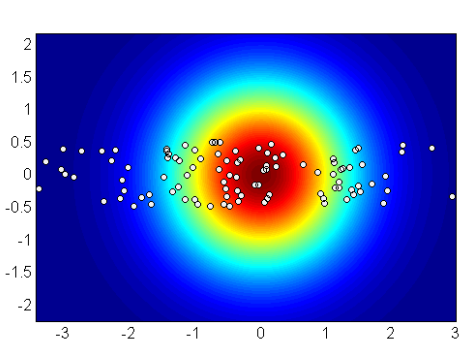


(b) Curvas de concordancia

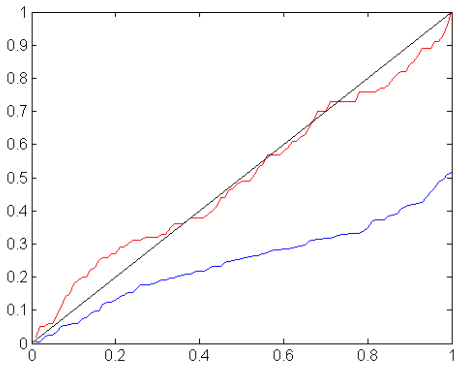
Figura 4.14: *Comparación entre una muestra estandarizada de tamaño 100 de un vector con componentes uniformes y la distribución normal estándar.*

Las muestras de los dos últimos ejemplos, Figuras 4.15 y 4.16, se han simulado de un vector compuesto por una coordenada normal con varianza $\sqrt{2}$ y otra uniforme entre -0.5 y 0.5, y de una normal estándar con un treinta por ciento de contaminación en torno al punto (2.5, 2.5)', respectivamente. En el primer ejemplo se observa cómo la concordancia de distribución con muestra recoge mejor la discrepancia, mientras que en el segundo

caso la situación es la opuesta.

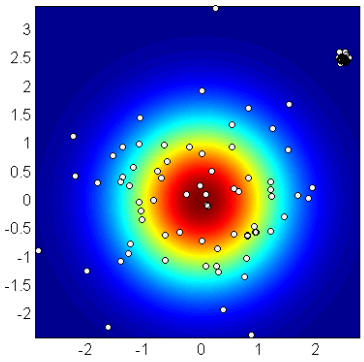


(a) Diagrama de dispersión y contornos esperados

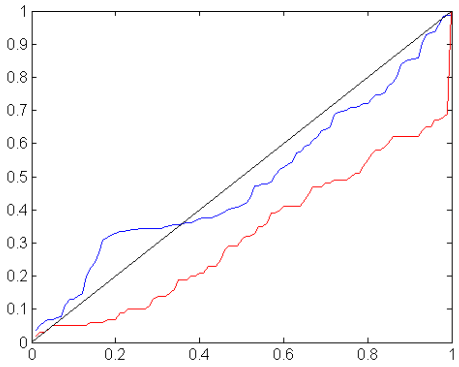


(b) Curvas de concordancia

Figura 4.15: *Comparación entre una muestra de tamaño 100 de un vector con componentes normal y uniforme y la distribución normal estándar.*



(a) Diagrama de dispersión y contornos esperados



(b) Curvas de concordancia

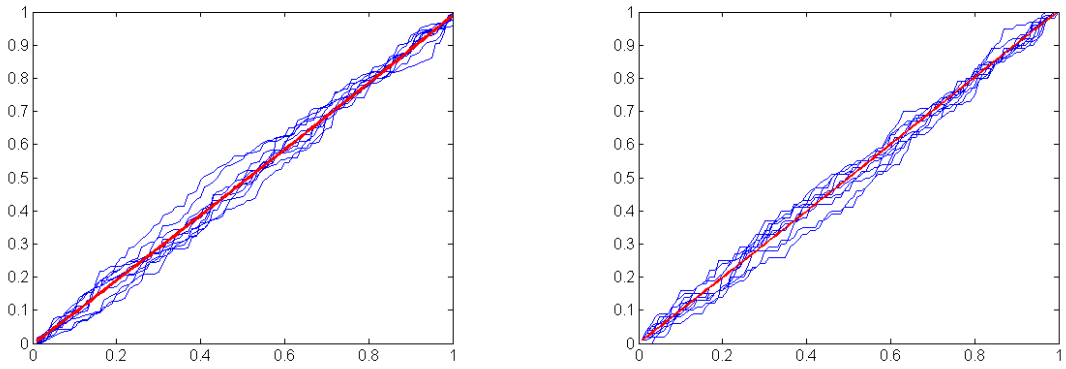
Figura 4.16: *Comparación entre una muestra de tamaño 100 de una normal contaminada en torno al punto (2.5,2.5)' y la distribución normal estándar.*

Las diferencias en el comportamiento de las curvas de concordancia en las Figuras 4.12, 4.13, 4.15 y 4.16 motiva el hecho de que el contraste tenga en cuenta ambas direcciones de análisis de discrepancia y no sólo una de ellas, ya que, en determinadas situaciones es posible que alguna de las medidas no detecte las diferencias existentes entre la muestra

y la distribución nula.

4.3.1. Estadístico del contraste

Al igual que en el contraste basado en la curva de escala, conviene medir las desviaciones de las curvas con respecto a la curva esperada si la distribución nula fuera cierta. La Figura 4.17 contiene una muestra de diez curvas de concordancia y una estimación de la curva esperada obtenida mediante simulación de 5000 muestras estandarizadas de la distribución normal estándar. Se observa cómo la curva esperada en ambos casos es la recta $y = x$.



(a) $CC_{F_o}(p)$ (b) $CC_n(p)$

Figura 4.17: *Muestra aleatoria de curvas de concordancia bajo H_0 y curva esperada.*

Observación 4.2 *Si la distribución nula es cierta, se tiene que, para cualquier $p \in [0,1]$, la variable aleatoria $nCC_n(p)$ sigue una distribución binomial de parámetros n y p, ya que la variable aleatoria dicotómica "pertenencia de x_i al conjunto C_p" es una variable aleatoria Bernouilli con parámetro la probabilidad de la región C_p que es igual a p.*

El estudio de la curva de concordancia de forma conjunta se podría llevar a cabo teniendo en cuenta que

$$Prob\left(nC_n(p) = x / nC_n(q) = y\right) = Prob\left(Binomial\left(n - y, \frac{p-q}{(1-q)}\right) = x - y\right),$$

donde $p \geq q$ y $x \geq y$.

Los percentiles de la muestra $P_n(x_i)$ pueden emplearse para construir una partición multivariante del espacio sobre la que aplicar la idea del contraste de bondad de ajuste chi-cuadrado, que podría mejorar los resultados del contraste en Watson (1957), Watson (1958) y Watson (1959) y sus posteriores versiones, en el que la partición del espacio se realiza mediante una elipse central rodeada por anillos elípticos. Con el uso de las funciones de profundidad se puede realizar una partición a priori más ajustada al conjunto de observaciones, en la que para cada región se asegura la pertenencia de un porcentaje de la muestra. No se tiene en cuenta esta posibilidad en este capítulo, ya que el hecho de realizar particiones presenta una cierta polémica, debido a que cada partición puede arrojar un resultado diferente y no se puede asegurar qué partición es la más adecuada.

Debido a que existe la posibilidad de que, en determinadas situaciones, alguna de las dos curvas no detecte las diferencias entre muestra y distribución, el estadístico de contraste que se propone integra a ambas curvas. Una posibilidad de integración es tomar el área entre las dos, pero no es adecuada ya que no siempre se tiene una de las curvas por encima de la recta $y = x$ y la otra por debajo, como en el ejemplo de la normal trasladada (Figura 4.11), en el que las dos curvas están por debajo de la esperada y la diferencia entre ellas es escasa. Para evitar ese problema se define el estadístico de contraste ACC como combinación lineal convexa de las discrepancias detectadas por ambas curvas, es decir, $ACC = \alpha ACC_n + (1-\alpha)ACC_{F_0}$ con $\alpha \in [0,1]$, donde ACC_n (ACC_{F_0}) representa la discrepancia detectada por la curva de concordancia de muestra con distribución (de distribución con muestra).

La función que se propone para cuantificar la aportación de cada curva es, como en el contraste basado en la curva de escala, de la forma

$$\text{Aportación} = \int_0^1 |Curva\,(p) - p|^k \,\mathrm{d}p,$$

es decir, para la concordancia de muestra con distribución se tiene que

$$ACC_n = \int_0^1 |CC_n\,(p) - p|^k \,\mathrm{d}p$$

y para la concordancia inversa

$$ACC_{F_0} = \int_0^1 |CC_{F_0}\,(p) - p|^k \,\mathrm{d}p,$$

donde $1 \leq k < \infty$. Se toma de nuevo el valor $k = 1$, es decir, el área entre la curva y su esperanza. Como a priori no se dispone información sobre qué curva detecta mejor las discrepancias, se toma para el peso de cada discrepancia el valor $\alpha = 0.5$. Por lo que se define el estadístico del contraste como

$$ACC = \frac{1}{2}\int_0^1 |CC_n(p) - p|\,\mathrm{d}p + \frac{1}{2}\int_0^1 |CC_{F_0}(p) - p|\,\mathrm{d}p.$$

Las funciones de cuantificación de cada curva pueden ser modificadas de modo que se pondere la variabilidad de las curvas para un determinado valor de p. La variabilidad de las curvas en los extremos es pequeña, mientras que para valores de p próximos a 0.5 es mayor. Una posible modificación es ponderar el valor absoluto por $p(1-p)$ o $\sqrt{p(1-p)}$, obteniéndose

$$ACC_n = \int_0^1 \frac{|CC_n(p) - p|}{\sqrt{p(1-p)}}\mathrm{d}p.$$

Para la obtención del estadístico sobre una muestra hay que calcular cuatro profundidades: $P_n(x_i)$, $P_n(X)$ con X distribuída según una normal, $P(x_i; F)$ y $P(X; F)$, donde F representa la función de distribución de la normal estándar y X tiene distribución F. Dependiendo de cuál sea la profundidad empleada para realizar las ordenaciones, se puede evitar el cálculo de las profundidades involucradas en $CC_n(p)$, reduciéndose el tiempo computacional para la estimación de esta concordancia. Las profundidades semiespacial, simplicial, de Oja, por proyecciones y L_1 verifican que $P(x; F) \geq P(y; F)$ si, y sólo si, $\|x\| \leq \|y\|$, y que $P(x; F) = P(y; F)$ si, y sólo si, $\|x\| = \|y\|$, donde F es la función de distribución normal estándar. Por lo que, para estas cinco profundidades, los valores de profundidad de un punto $x \in \mathbb{R}^d$ con respecto a la normal estándar se relacionan con la norma euclídea del punto x mediante una función monótona, lo que, unido a que la norma al cuadrado bajo F sigue una distribución χ^2_d, justifica que sea equivalente el comparar los valores de profundidad y sus cuantiles con comparar las normas y sus cuantiles.

4.3.2. Valores críticos

A pesar de que sea factible el estudio de la distribución de la concordancia de muestra con hipótesis nula, los valores críticos que se introducen a continuación se han estimado

mediante simulación, debido a que el estudio teórico de la distribución de la concordancia de distribución nula con muestra resulta muy complejo.

La tabla 4.19 contiene los valores críticos estimados para contrastes con niveles de significación 0.1, 0.05 y 0.01, para todas las profundidades empleadas en el contraste de bondad de ajuste anterior y para tamaños muestrales 50 y 100.

		Percentil		
Profundidad	Tamaño muestral	0.90	0.95	0.99
Semiespacio	50	0.0540	0.0605	0.0730
	100	0.0370	0.0410	0.0500
Simplicial	50	0.1355	0.1430	0.1595
	100	0.0840	0.0900	0.1020
Oja	50	0.0510	0.0570	0.0720
	100	0.0350	0.0395	0.0500
Proyecciones	50	0.0490	0.0555	0.0700
	100	0.0340	0.0385	0.0475
L_1	50	0.0555	0.0625	0.0745
	100	0.0370	0.0420	0.0520
Bandas	50	0.0980	0.1060	0.1230
	100	0.0580	0.0635	0.0730
Bandas modificada	50	0.0504	0.0573	0.0722
	100	0.0367	0.0410	0.0513

Tabla 4.19: *Valores críticos para el contraste de bondad de ajuste basado en las curvas de concordancia para distribución nula normal.*

Para distribuciones nulas uniforme y exponencial, no se ha estudiado si existe alguna propiedad que permita evitar el cálculo de las profundidades con respecto a la función de distribución nula. En el caso de normalidad es posible utilizar los cuantiles de la distribución chi-cuadrado para obtener estas profundidades. Por lo tanto, para estas dos distribuciones nulas, es necesario realizar estimaciones de las profundidades con respecto a la distribución nula. Esto se lleva a cabo mediante muestras generadas a partir de la nula de tamaño 10000. Además, debido al elevado coste computacional de estos cálculos para algunas funciones de profundidad, solo se aplica el algoritmo a las profundidades

por proyecciones, bandas y bandas modificada. Los valores críticos estimados con 10000 muestras están recogidos en la Tabla 4.20.

		Uniforme			Exponencial		
		Percentil			Percentil		
Profundidad	Tamaño muestral	0.90	0.95	0.99	0.90	0.95	0.99
Proyecciones	50	0.0717	0.0816	0.1028	0.0577	0.0664	0.0853
	100	0.0516	0.0594	0.0770	0.0413	0.0473	0.0606
Bandas	50	0.1043	0.1145	0.1379	0.1208	0.1326	0.1579
	100	0.0660	0.0741	0.0910	0.0750	0.0847	0.1038
Bandas modificada	50	0.0755	0.0878	0.1156	0.0699	0.0806	0.1033
	100	0.0535	0.0626	0.0804	0.0502	0.0575	0.0735

Tabla 4.20: *Valores críticos para el contraste de bondad de ajuste de la curva de concordancia para distribución nula uniforme y exponencial.*

4.3.3. Potencia del contraste

El estudio de la potencia del contraste se realiza sobre las mismas hipótesis alternativas que en el contraste de la curva de escala, construyéndose el mismo índice de ordenación de profundidades para medir cuál se comporta mejor. Se comienza con el contraste de la hipótesis nula de normalidad.

4.3.3.1. Potencia para distribución nula normal

Las Tablas 4.21 y 4.22 contienen las potencias estimadas para las alternativas del primer grupo, formadas por vectores bidimensionales de coordenadas independientes. En este contraste se observa una mayor homogeneidad entre las distintas profundidades, salvo en los casos en que la distribución beta está en alguna de las dos coordenadas, donde las profundidades simplicial, por bandas y por bandas modificada no son capaces de determinar diferencia alguna con respecto a la normal.

Sobre muestras pequeñas la profundidad que sustancialmente mejor se comporta para este grupo es la profundidad de Oja (Tabla 4.23), que conserva su posición como mejor

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
Exponencial	50	93	73	96	94	91	90	88
	100	100	97	100	100	100	100	99
Lognormal	50	99	98	100	100	100	100	100
	100	100	100	100	100	100	100	100
Gamma(5,1)	50	20	17	24	22	22	21	23
	100	29	25	36	36	29	36	33
chi-cuadrado(5)	50	42	31	52	49	45	45	42
	100	67	55	79	76	67	72	66
chi-cuadrado(15)	50	12	11	16	14	12	14	15
	100	16	15	22	22	17	23	21
t(2)	50	99	99	99	98	98	99	99
	100	100	100	100	100	100	100	100
t(5)	50	51	58	54	53	57	61	60
	100	77	82	80	78	83	87	84
Logística(0,1)	50	27	34	31	27	32	37	37
	100	45	53	50	46	50	59	59
Beta(1,1)	50	82	0	74	70	66	1	0
	100	100	21	99	99	99	69	46
Beta(1,2)	50	40	0	20	17	18	0	0
	100	80	1	48	43	55	11	3
Beta(2,2)	50	32	0	22	22	16	0	0
	100	66	0	58	61	54	5	2

Tabla 4.21: *Potencia del contraste de bondad de ajuste basado en las curvas de concordancia con distribución nula normal, para vectores bidimensionales cuyas componentes son independientes e igualmente distribuidas.*

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
Normal(0,1) y Exponencial	50	46	38	62	63	49	52	54
	100	76	62	87	88	76	85	74
Normal(0,1) y chi-cuadrado(5)	50	17	15	24	25	22	21	24
	100	26	23	36	38	29	36	30
Normal(0,1) y t(5)	50	24	28	25	27	25	31	31
	100	38	43	40	39	37	51	50
Normal(0,1) y Beta(1,1)	50	23	0	17	18	15	0	0
	100	58	1	48	48	47	5	2
Normal(0,1) y Beta(1,2)	50	12	1	8	5	7	1	1
	100	25	0	15	11	14	2	1

Tabla 4.22: *Potencia del contraste de bondad de ajuste basado en las curvas de concordancia con distribución nula normal, para vectores bidimensionales cuyas componentes son independientes y poseen distribuciones diferentes.*

también para muestras grandes, aunque las diferencias con el resto en ese caso sean menores. En segundo lugar se encuentra la profundidad por proyecciones que, si bien ocupa el tercer lugar en muestras de tamaño 100, la diferencia con la segunda (profundidad por bandas) es pequeña, mientras que en muestras de 50 observaciones ocupa el segundo lugar con una considerable distancia sobre la tercera.

	Profundidad de ordenación						
n	PSem	PS	PO	PP	PL_1	PB	PBM
50	4.16	5.72	2.59	3.41	4.16	4.06	3.91
100	3.91	5.94	3.00	3.25	4.13	3.19	4.59

Tabla 4.23: *Índice de rangos para el contraste basado en las curvas de concordancia con distribución nula normal y el grupo 1 de alternativas.*

En la Tabla 4.24, correspondiente a las mixturas de normales, se puede observar que, cuando las normales que componen la mezcla sólo se diferencian en la media, las profundidades simplicial, por bandas y por bandas modificada se comportan de nuevo muy por debajo del resto, mientras que cuando también se tienen diferencias en la matriz de covarianzas llegan incluso a ser mejores que las restantes. De forma global sobre este grupo de alternativas, Tabla 4.25, se tiene que la mejor profundidad con diferencia es la semiespacial.

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
Mixtura Normal (2,0,0)	50	13	0	10	9	8	0	1
	100	31	0	25	26	20	1	2
Mixtura Normal (4,0,0)	50	73	0	60	59	64	2	1
	100	99	28	97	97	98	69	41
Mixtura Normal (2,0.9,0)	50	34	29	30	30	19	31	22
	100	49	47	43	43	28	52	38
Mixtura Normal (0.5,0.9,0)	50	34	39	38	37	37	40	31
	100	59	67	57	64	58	64	52
Mixtura Normal (0.5,0.9,-0.9)	50	78	85	76	64	72	77	53
	100	97	99	96	91	93	98	93

Tabla 4.24: *Potencia del contraste de bondad de ajuste basado en las curvas de concordancia con distribución nula normal, para mixturas de normales bidimensionales.*

	Profundidad de ordenación						
n	PSem	PS	PO	PP	PL_1	PB	PBM
50	2.20	4.30	3.10	4.20	4.50	3.50	6.20
100	2.20	3.80	4.20	3.90	4.70	3.30	5.90

Tabla 4.25: *Índice de rangos para el contraste basado en las curvas de concordancia con distribución nula normal y el grupo 2 de alternativas.*

Para el grupo de distribuciones alternativas esféricas (Tabla 4.26) se obtienen de forma global mejores resultados para las distribuciones de Pearson que para las esféricas en las que se especifica la distribución del radio. De nuevo, las profundidades simplicial, por

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
PearsonII(0)	50	92	0	85	85	83	4	3
	100	100	40	100	100	100	87	76
PearsonII(1)	50	42	0	33	36	24	0	0
	100	88	1	80	83	72	14	6
PearsonVII(2)	50	99	99	99	99	99	99	99
	100	100	100	100	100	100	100	100
PearsonVII(3)	50	78	82	80	76	80	82	79
	100	96	97	97	96	95	97	97
PearsonVII(5)	50	34	39	35	36	39	42	39
	100	55	63	58	56	57	62	62
Esférica(Exponencial)	50	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100
Esférica(Gamma(5,1))	50	23	1	18	17	14	1	1
	100	50	2	47	43	44	7	3
Esférica(Beta(1,1))	50	33	19	34	38	27	18	15
	100	75	28	78	82	67	38	36
Esférica(Beta(1,2))	50	79	83	82	83	82	79	78
	100	98	98	99	99	98	98	98
Esférica(Beta(2,2))	50	20	0	16	16	10	0	0
	100	47	0	45	43	34	2	0

Tabla 4.26: *Potencia del contraste de bondad de ajuste basado en las curvas de concordancia con distribución nula normal, para distribuciones de Pearson y esféricas.*

bandas y por bandas modificada encuentran problemas en varios de los ejemplos. El índice (Tabla 4.27) en este grupo muestra que la mejor profundidad es la de Oja, seguida de la de proyecciones y la simplicial, que tienen un comportamiento global muy parecido.

Para el cuarto grupo (Tabla 4.28) se sigue detectando el problema de potencia en los casos simplicial, por bandas y por bandas modificada, si bien si se observa cómo, a medida que aumenta el valor de la correlación, se tiene una mayor potencia. En esta ocasión, según puede observarse en la Tabla 4.29, la mejor profundidad es por proyecciones, seguida de Oja y de semiespacial, que experimenta una notable mejora cuando el tamaño muestral es igual a 100.

	Profundidad de ordenación						
n	PSem	PS	PO	PP	PL_1	PB	PBM
50	3.35	4.40	3.30	3.25	3.80	4.50	5.40
100	3.40	5.10	2.75	3.35	4.25	4.30	4.85

Tabla 4.27: *Índice de rangos para el contraste basado en las curvas de concordancia con distribución nula normal y el grupo 3 de alternativas.*

		Profundidad de ordenación						
Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM
0.2	50	4	4	7	6	5	6	6
	100	6	6	6	6	5	5	4
0.4	50	6	6	9	10	7	8	6
	100	9	6	12	12	8	8	8
0.6	50	13	7	14	20	11	9	10
	100	16	8	19	28	12	11	12
0.8	50	22	9	25	38	20	10	13
	100	36	9	34	69	23	12	17
1	50	37	9	42	76	29	12	17
	100	71	9	67	99	45	18	24

Tabla 4.28: *Potencia del contraste de bondad de ajuste basado en las curvas de concordancia con distribución nula normal, para distribuciones con correlación radial/angular.*

	Profundidad de ordenación						
n	PSem	PS	PO	PP	PL_1	PB	PBM
50	4.30	6.70	1.80	1.40	4.20	4.80	4.80
100	2.50	6.10	2.40	1.40	4.60	5.70	5.30

Tabla 4.29: *Índice de rangos para el contraste basado en las curvas de concordancia con distribución nula normal y el grupo 4 de alternativas.*

Finalmente, sobre las 36 alternativas (Tabla 4.30), se tiene, como en el contraste de la curva de escala, que la ordenación que mejor se comporta es la de Oja, seguida de proyecciones, apareciendo en tercer lugar la del semiespacio.

	Profundidad de ordenación						
n	PSem	PS	PO	PP	PL_1	PB	PBM
50	3.68	5.29	2.75	3.19	4.11	4.21	4.76
100	3.33	5.43	3.01	3.11	4.31	3.86	4.94

Tabla 4.30: *Índice de rangos global para el contraste basado en las curvas de concordancia con distribución nula normal*

4.3.3.2. Potencia para distribución nula uniforme

Como ya se ha comentado en la sección del valor crítico, el contraste de la curva de concordancia para distribuciones nulas uniforme y exponencial se ha aplicado sobre las profundidades por proyecciones, bandas y bandas modificada. A continuación se muestran los porcentajes de rechazo obtenidos, para cada una de las veinte alternativas, a partir de 1000 simulaciones.

La Tabla 4.31 contiene los porcentajes de rechazo estimados para vectores con coordenadas distribuidas según la distribución beta, para la uniforme en la circunferencia unidad y para la normal con y sin truncamiento. Se observa que existe una elevada heterogeneidad dentro de cada grupo de distribuciones. Así, por ejemplo, para la beta con parámetros menores que uno, los mejores resultados corresponden a la profundidad por proyecciones y a la de bandas modificada, mientras que si son mayores que uno, se dan para la de proyecciones y la de bandas. Para la distribución normal sin truncamiento, todos los porcentajes están en el entorno del 100 %. Para la normal con truncamiento, el comportamiento de la profundidad por bandas está por encima que el de las demás.

Para las mixturas de uniformes (Tabla 4.32), la mejor es la profundidad por proyecciones, seguida de la de bandas modificada. Mientras que para las uniformes en cuadrados recortados y la distribución de Pearson, la profundidad por bandas es superior al resto.

De forma global (Tabla 4.33), la que presenta tanto un porcentaje de rechazo medio más elevado, como un índice de rangos medio más bajo, es la profundidad por bandas, seguida por la de proyecciones.

		Profundidad		
Distribución	n	PP	PB	PBM
Beta(0.8,0.8)	50	30	2	36
	100	48	21	56
Beta(0.9,0.9)	50	12	1	13
	100	17	4	19
Beta(1.15,1.15)	50	6	16	4
	100	15	29	9
Beta(1.3,1.3)	50	14	30	7
	100	41	65	32
Unif. Circunferencia	50	20	35	13
	100	66	78	62
Normal	50	100	100	99
	100	100	100	100
Normal circ. (1)	50	43	60	34
	100	91	95	89
Normal circ. (1.5)	50	71	85	64
	100	100	100	99
Normal circ. (2)	50	92	98	87
	100	100	100	100

Tabla 4.31: *Potencia del contraste de bondad de ajuste basado en las curvas de concordancia con distribución nula uniforme y alternativas beta, uniforme en circunferencia y normal.*

4.3.4. Potencia para distribución nula exponencial

La Tabla 4.34 contiene los porcentajes de rechazo para distribuciones alternativas normal, lognormal, chi-cuadrado y gamma. Para la chi-cuadrado con 4 o más grados de libertad, así como para la normal y la gamma, estos porcentajes son, para las tres funciones de profundidad, iguales a 100 o a valores muy próximos. Para la distribución lognormal las tres profundidades obtienen resultados similares que se sitúan en torno al 55 %. Por último, para el valor absoluto de la distribución normal, se tiene que la profundidad por proyecciones obtiene una potencia menor que las otras dos.

Los resultados para las alternativas distribuidas según la distribución Weibull y para

		Profundidad		
Distribución	n	PP	PB	PBM
Mixt. Unif. (0.1,0.5)	50	5	6	4
	100	7	5	4
Mixt. Unif. (0.25,0.5)	50	18	13	4
	100	33	14	5
Mixt. Unif. (0.1,0.25)	50	13	3	15
	100	21	3	18
Mixt. Unif. (0.25,0.25)	50	74	16	34
	100	91	28	53
Cuad. Recort (0.2)	50	4	6	4
	100	4	9	4
Cuad. Recort (0.4)	50	6	13	3
	100	14	22	8
Cuad. Recort (0.6)	50	16	31	7
	100	39	53	22
Cuad. Recort (0.8)	50	58	59	16
	100	93	95	44
Cuad. Recort (1)	50	88	90	34
	100	100	100	88
Pearson II (0)	50	18	35	13
	100	68	75	60
Pearson II (1)	50	79	90	72
	100	100	100	100

Tabla 4.32: *Potencia del contraste de bondad de ajuste basado en las curvas de concordancia con distribución nula uniforme y alternativas mixtura de uniforme, uniforme en cuadrados recortados y Pearson II.*

		Profundidad de ordenación		
	n	PP	PB	PBM
Potencia media	50	38.35	39.45	28.15
	100	57.40	54.80	48.60
Rango medio	50	1.90	1.48	2.63
	50	1.78	1.70	2.53

Tabla 4.33: *Porcentaje medio de rechazo e índice de rangos para el contraste de bondad de ajuste de la curva de concordancia sobre distribución nula uniforme.*

		Profundidad		
Distribución	n	PP	PB	PBM
Normal	50	100	100	100
	100	100	100	100
|Normal|	50	38	61	50
	100	59	84	73
Lognormal	50	54	57	55
	100	81	82	80
Chi-cuadrado (1)	50	46	35	97
	100	88	98	100
Chi-cuadrado (3)	50	53	85	77
	100	77	98	96
Chi-cuadrado (4)	50	94	100	100
	100	100	100	100
Chi-cuadrado (5)	50	100	100	100
	100	100	100	100
Chi-cuadrado (10)	50	100	100	100
	100	100	100	100
Gamma (5,1)	50	100	100	100
	100	100	100	100

Tabla 4.34: *Potencia del contraste de bondad de ajuste basado en las curvas de concordancia con distribución nula exponencial y alternativas normal, lognormal, gamma y chi-cuadrado.*

las mixturas de exponenciales se encuentran en la Tabla 4.35. Se observa cómo, de forma generalizada para todos los valores de los parámetros, la profundidad por bandas modificada mejora sustancialmente los porcentajes de proyecciones y bandas.

Como sugieren los resultados de la tablas anteriores, el porcentaje de rechazo medio más alto se da para la profundidad por bandas modificada (Tabla 4.36), que alcanza para 50 observaciones el 60 %. Tras ésta se sitúa la profundidad por bandas con un 53 % de rechazo. Según el índice de posiciones para cada alternativa la segunda posición estaría ocupada por la profundidad por proyecciones.

		Profundidad		
Distribución	n	PP	PB	PBM
Weibull (1,0.5)	50	97	94	100
	100	100	100	100
Weibull (1,0.75)	50	12	3	59
	100	32	38	92
Weibull (1,0.9)	50	3	1	9
	100	5	0	22
Weibull (1,1.1)	50	16	23	19
	100	13	36	30
Weibull (1,1.3)	50	57	84	80
	100	79	98	95
Weibull (1,1.7)	50	99	100	100
	100	100	100	100
Mixt. Expo. (0.3,0.2)	50	16	1	35
	100	34	12	73
Mixt. Expo. (0.2,0.2)	50	6	0	21
	100	14	2	45
Mixt. Expo. (0.1,0.2)	50	3	1	7
	100	6	1	16
Mixt. Expo. (0.3,0.5)	50	5	2	3
	100	4	2	6
Mixt. Expo. (0.2,0.5)	50	5	3	4
	100	5	2	6
Mixt. Expo. (0.1,0.5)	50	4	4	5
	100	5	4	3

Tabla 4.35: *Potencia del contraste de bondad de ajuste basado en las curvas de concordancia con distribución nula exponencial y alternativas Weibull y mixtura de exponenciales.*

		Profundidad de ordenación		
	n	PP	PB	PBM
Potencia media	50	49.81	53.10	60.53
	100	60.05	63.86	71.90
Rango medio	50	2.26	2.17	1.57
	100	2.24	2.05	1.71

Tabla 4.36: *Porcentaje medio de rechazo e índice de rangos para el contraste de bondad de ajuste basado en las curvas de concordancia sobre distribución nula exponencial.*

4.4. Contraste basado en similaridades

En Liu et al. (1999) se propone el uso de los *dd*-plots para comparar dos muestras, dos distribuciones o una muestra con una distribución. Los *dd*-plots son diagramas de dispersión en los que cada coordenada representa la profundidad P respecto a un determinado conjunto de datos o una determinada función de distribución. Por ejemplo, para realizar comparaciones entre una muestra $x_1, x_2, \ldots, x_n$ y otra $y_1, y_2, \ldots, y_m$, se representan todos los puntos de coordenadas

$$\left(P_{F_n}(x), P_{G_m}(x)\right), x \in \mathbb{X} \cup \mathbb{Y},$$

donde $\mathbb{X}$ e $\mathbb{Y}$ representan el conjunto de las observaciones de cada muestra, y F y G las distribuciones de X e Y, respectivamente. Para comparar dos distribuciones F y G se realiza el diagrama de puntos de

$$\left(P_F(x), P_G(x)\right), x \in \mathbb{R}^d.$$

Finalmente si se desea comparar una muestra $x_1, x_2, \ldots, x_n$ de distribución desconocida F con una distribución G se representan los puntos

$$\left(P_{F_n}(x), P_G(x)\right), x \in \{x_1, x_2, \ldots, x_n\}.$$

Hasta la fecha no se ha propuesto ningún estadístico basado en la idea del *dd*-plot para contrastar si una muestra sigue una determinada distribución. Es por lo tanto un método gráfico del tipo de los gráficos cuantil-cuantil. Los *dd*-plots, según puede verse en Liu et al. (1999), pueden detectar cambios tanto en la media como en la varianza, la asimetría y la curtosis.

El objetivo de esta sección no es directamente el de cuantificar la discrepancia de un *dd*-plot, sino cuantificar las discrepancias que ofrecen n *dd*-plots. Con las funciones de profundidad se obtiene una medida de la proximidad al centro que hace posible la comparación por medio de estos diagramas de dispersión. Sin embargo, con las funciones de similaridad introducidas en el Capítulo 2, se obtiene, para cada observación de una muestra de n observaciones, los $n-1$ valores de la proximidad de los restantes puntos de la muestra, más su posición con respecto al centro (su profundidad).

En esta sección, las similaridades sobre las que se aplica el contraste son la de Oja, por proyecciones, por bandas y por bandas modificada. No se tiene en cuenta la similaridad simplicial ya que no se dispone de ningún algoritmo que permita su uso para el cálculo tanto de valores críticos como de potencia.

Según se definieron estas cuatro similaridades en el Capítulo 2, se tiene una primera diferencia entre ellas: las profundidades de Oja y por proyecciones entre dos puntos iguales es igual a 1 ($SO(x,x;F) = 1$ y $SP(x,x;F) = 1$), mientras que para las similaridades por bandas y por bandas modificada es igual a la profundidad del punto ($SB(x,x;F) = PB(x;F)$ y $SBM(x,x;F) = PB(x;F)$). Para establecer una definición adecuada es necesario homogeneizar estas diferencias.

Definición 4.3 *Dadas una distribución d-dimensional F y una muestra aleatoria $x_1, x_2, \ldots, x_n$, se define la matriz de similaridad de Oja estandarizada ($MSOE_F$) como*

$$MSOE_F = \begin{pmatrix} 1 & SO_F(x_1,x_2) & \cdots & SO_F(x_1,x_n) \\ SO_F(x_2,x_1) & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & SO_F(x_{n-1},x_n) \\ SO_F(x_n,x_1) & \cdots & SO_F(x_n,x_{n-1}) & 1 \end{pmatrix}$$

y la matriz de similaridad de Oja (MSO_F) como $MSO_F = D^{1/2} \cdot MSOE_F \cdot D^{1/2}$, donde

$$D = \begin{pmatrix} PO_F(x_1) & 0 & \cdots & 0 \\ 0 & PO_F(x_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & PO_F(x_n) \end{pmatrix}.$$

De forma análoga se definen las matrices para la similaridad y profundidad por proyecciones.

Definición 4.4 *Dadas una distribución d-dimensional F y una muestra aleatoria $x_1, x_2,$*

$\ldots, x_n$, se define la matriz de similaridad por proyecciones estandarizada ($MSPE_F$) como

$$MSPE_F = \begin{pmatrix} 1 & SP_F(x_1,x_2) & \cdots & SP_F(x_1,x_n) \\ SP_F(x_2,x_1) & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & SP_F(x_{n-1},x_n) \\ SP_F(x_n,x_1) & \cdots & SP_F(x_n,x_{n-1}) & 1 \end{pmatrix}$$

y la matriz de similaridad por proyecciones (MSP_F) como $MSP_F = D^{1/2} \cdot MSPE_F \cdot D^{1/2}$, donde

$$D = \begin{pmatrix} PP_F(x_1) & 0 & \cdots & 0 \\ 0 & PP_F(x_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & PP_F(x_n) \end{pmatrix}.$$

Las matrices para las similaridades por bandas y por bandas modificada se obtienen de forma inversa.

Definición 4.5 *Dadas una distribución d-dimensional F y una muestra aleatoria x_1, x_2, $\ldots, x_n$, se define la matriz de similaridad por bandas (MSB_F) como*

$$MSB_F = \begin{pmatrix} SB_F(x_1,x_1) & SB_F(x_1,x_2) & \cdots & SB_F(x_1,x_n) \\ SB_F(x_2,x_1) & SB_F(x_2,x_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & SB_F(x_{n-1},x_n) \\ SB_F(x_n,x_1) & \cdots & SB_F(x_n,x_{n-1}) & SB_F(x_n,x_n) \end{pmatrix}$$

y la matriz de similaridad por bandas estandarizada ($MSBE_F$) como $MSBE_F = D^{-1/2} \cdot MSB_F \cdot D^{-1/2}$, donde

$$D = \begin{pmatrix} PB_F(x_1) & 0 & \cdots & 0 \\ 0 & PB_F(x_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & PB_F(x_n) \end{pmatrix}.$$

Definición 4.6 *Dadas una distribución d-dimensional F y una muestra aleatoria $x_1, x_2, \ldots, x_n$, se define la matriz de similaridad por bandas modificada ($MSBM_F$) como*

$$MSBM_F = \begin{pmatrix} SBM_F(x_1,x_1) & SBM_F(x_1,x_2) & \cdots & SBM_F(x_1,x_n) \\ SBM_F(x_2,x_1) & SBM_F(x_2,x_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & SBM_F(x_{n-1},x_n) \\ SBM_F(x_n,x_1) & \cdots & SBM_F(x_n,x_{n-1}) & SBM_F(x_n,x_n) \end{pmatrix}$$

y la matriz de similaridad por bandas modificada estandarizada como $MSBME_F = D^{-1/2} \cdot MSBM_F \cdot D^{-1/2}$, donde

$$D = \begin{pmatrix} PBM_F(x_1) & 0 & \cdots & 0 \\ 0 & PBM_F(x_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & PBM_F(x_n) \end{pmatrix}.$$

Gracias a estas matrices de similaridad (MS) y de similaridad estandarizada (MSE) es posible analizar de manera gráfica, a través de diagramas de dispersión, las discrepancias existentes entre muestra y distribución. En estos gráficos similaridad-similaridad o *ss*-plot, se representan los pares de puntos $(MS_F(i,j), MS_{F_n}(i,j))$ (para las matrices sin estandarizar) o $(MSE_F(i,j), MSE_{F_n}(i,j))$ (para las matrices estandarizadas) donde $MS_F(i,j)$ ($MS_{F_n}(i,j)$) es el elemento j-ésimo de la fila i-ésima de la matriz MS_F (MS_{F_n}), MS_F es la matriz de similaridad teórica bajo F y MS_{F_n} es la matriz de similaridad muestral o bajo F_n.

Las Figuras 4.18 y 4.19 contienen, para las matrices de similaridades por bandas MSB y $MSBE$, los *ss*-plots de muestras simuladas a partir de distribuciones normal, uniforme, exponencial y normal contaminada, comparadas con la función de distribución teórica normal bivariante. Puede observarse cómo en los gráficos en los que las muestras se generan a partir de la distribución normal (Figuras 4.18(a) y 4.19(a)), los puntos se disponen de forma equilibrada por encima y por debajo de la bisectriz del primer cuadrante que representa la línea esperada. No sucede lo mismo en el resto de las muestras. Cuando la muestra es uniforme los puntos presentan una forma no lineal, estando su mayoría por

encima de la recta. Hecho que se cumple con más claridad en la Figura 4.18(b) que en la 4.19(b). Para muestras de exponenciales independientes (Figuras 4.18(c) y 4.19(c)), la mayoría de los puntos se sitúa por debajo de la recta, pudiéndose observar además que la variabilidad aumenta significativamente con respecto a los ejemplos anteriores, dándose este aumento de forma más notable para la matriz de similaridad estandarizada. Por último, para la muestra generada a partir de una normal contaminada, se tiene que en el caso de la matriz MSB (Figura 4.18(d)), la disposición de los puntos es semejante a la de la muestra normal, si bien está ligeramente sesgada por encima de la recta. Si se emplea la matriz $MSBE$ se detecta de forma más clara la desviación con respecto a la teórica debido al aumento de la variabilidad.

4.4.1. Estadístico del contraste

En esta sección se introduce el estadístico que cuantifica las desviaciones entre muestra y distribución de los *ss*-plots. Éste depende de si la matriz empleada es la estandarizada o la que contiene los valores de la profundidad en la diagonal principal y consiste en el promedio de los valores absolutos de las desviaciones entre las similaridades muestrales y las esperadas o teóricas bajo la hipótesis nula de normalidad.

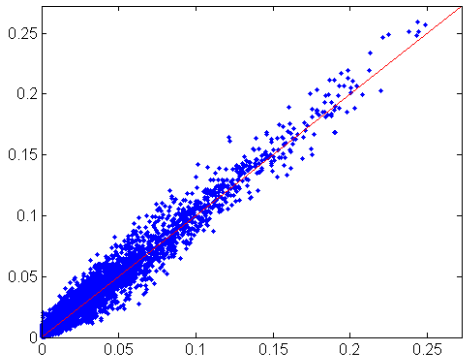
Así se tiene que, si la matriz empleada es la de similaridad (las profundidades en la diagonal principal), se define el estadístico DS como

$$DS = \left(\frac{n\,(n+1)}{2}\right)^{-1} \sum_{i=1}^{n} \sum_{j=i}^{n} \left| MS_{F_n}(i,j) - MS_F(i,j) \right|,$$

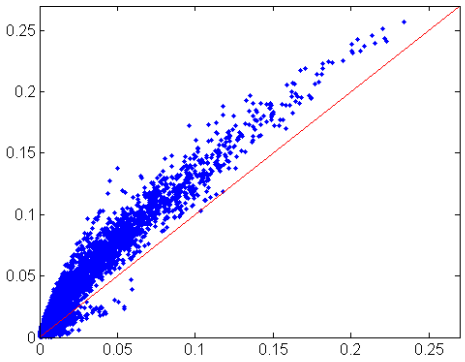
y que, si se emplea la matriz estandarizada, se define el estadístico DSE como

$$DSE = \left(\frac{n\,(n-1)}{2}\right)^{-1} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left| MSE_{F_n}(i,j) - MSE_F(i,j) \right|.$$

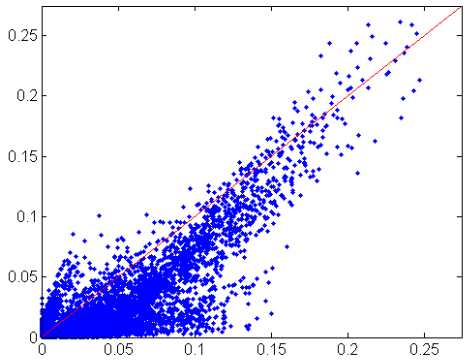
Para ilustrar cómo se comportan estos estadísticos ante varias muestras simuladas y comparando con la distribución normal, se introducen las Figuras 4.20 a 4.27. Las muestras para las que se obtienen estas figuras se han simulado de cuatro distribuciones: normal, exponencial, uniforme y normal contaminada. Para cada una de las cuatro muestras se tienen dos figuras: en una se toma como punto fijo uno próximo a la media de



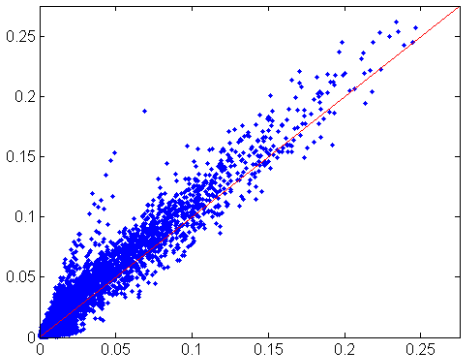
(a) Muestra normal



(b) Muestra uniforme



(c) Muestra exponencial

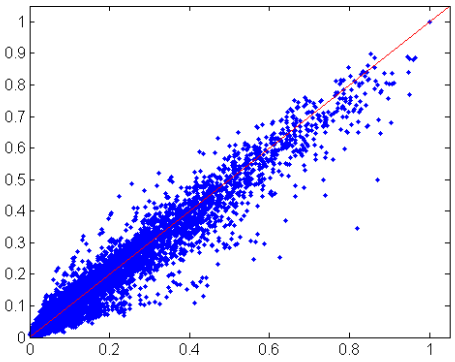


(d) Muestra normal contaminada

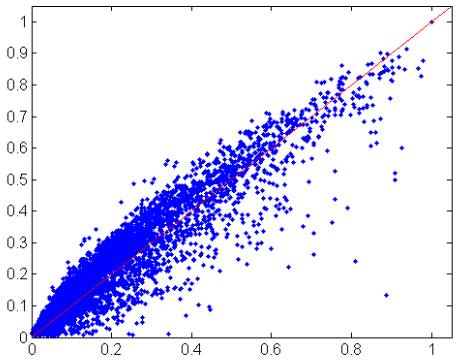
Figura 4.18: *Diagramas de puntos de la matriz de similaridades por bandas teórica (eje X) frente a la muestral (eje Y).*

la muestra, y en la otra un punto externo alejado de ésta. Cada una de las ocho figuras está compuesta por cuatro gráficos que representan superficies. Los dos primeros muestran la superficie de la similaridad muestral y la teórica, y los últimos las superficies de los valores absolutos de las diferencias entre las similaridades muestrales y teóricas y entre las similaridades estandarizadas muestrales y teóricas. La similaridad empleada para su obtención es la similaridad por bandas.

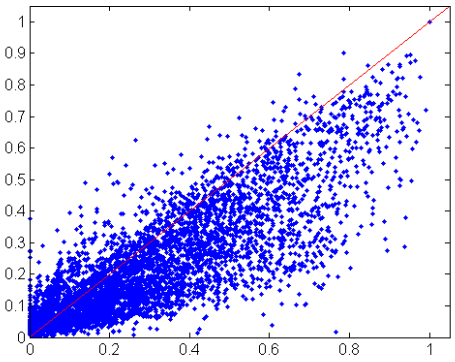
Las Figuras 4.20 y 4.21 corresponden a la muestra de distribución normal. Puede observarse la elevada similitud existente entre las similaridades muestrales y las teóricas tanto para el punto central (4.20(a) y 4.20(b)) como para el punto externo (4.21(a)



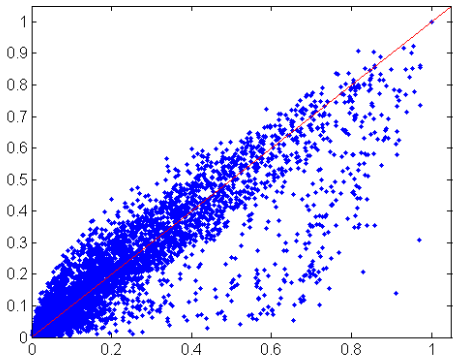
(a) Muestra normal



(b) Muestra uniforme



(c) Muestra exponencial

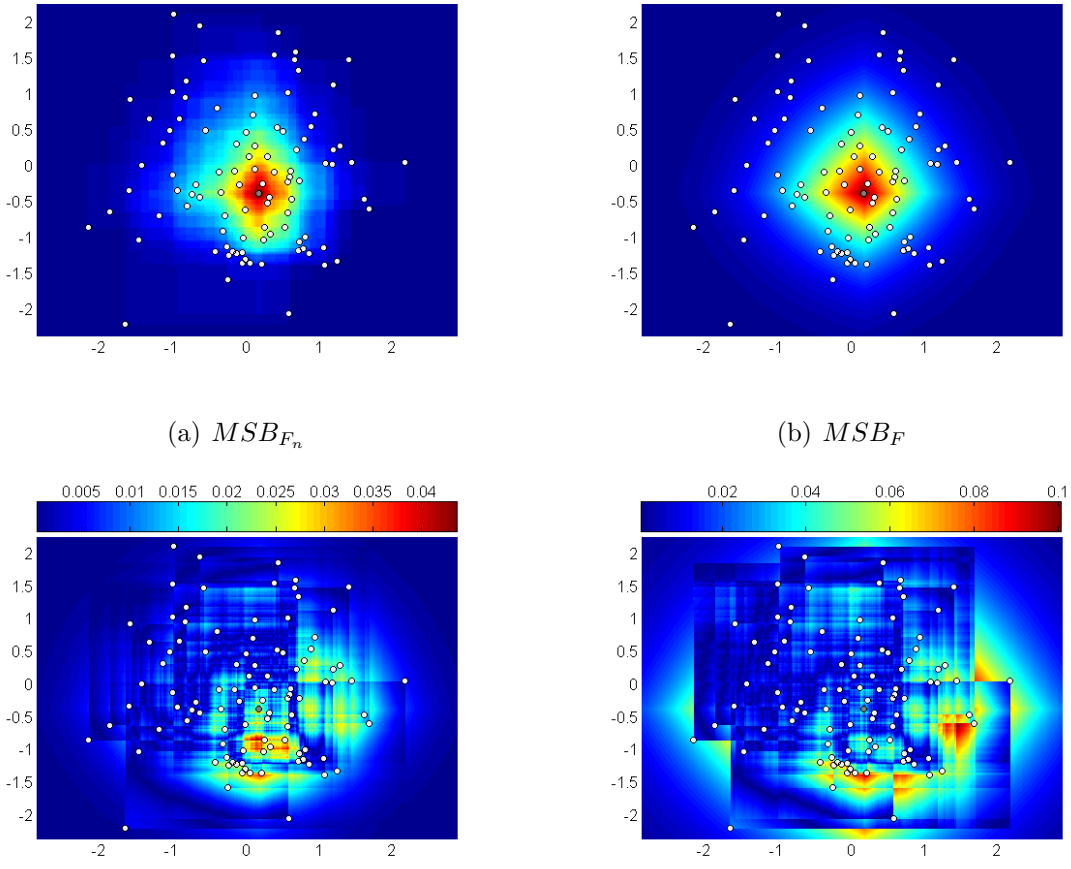


(d) Muestra normal contaminada

Figura 4.19: *Diagramas de puntos de la matriz de similaridades por bandas estandarizada teórica (eje X) frente a la muestral (eje Y).*

y 4.21(b)). Debido a esta similitud se tiene que las superficies de $|MSB_F - MSB_{F_n}|$ y $|MSBE_F - MSBE_{F_n}|$ no presentan amplias zonas con elevados valores de esta diferencia, siendo, en el primer caso todas menores que 0.04 y en el segundo menores que 0.1, ya sea el punto fijo central o externo.

Las superficies para la muestra con coordenadas independientres de distribución exponencial se encuentran en las Figuras 4.22 y 4.23. Puede observarse cómo disminuye respecto al ejemplo anterior la similitud entre las similaridades muestrales y las teóricas tanto para el punto central (4.22(a) y 4.22(b)) como para el externo (4.23(a) y 4.23(b)), teniéndose una mayor diferencia para el central, debido a que la teórica presenta una



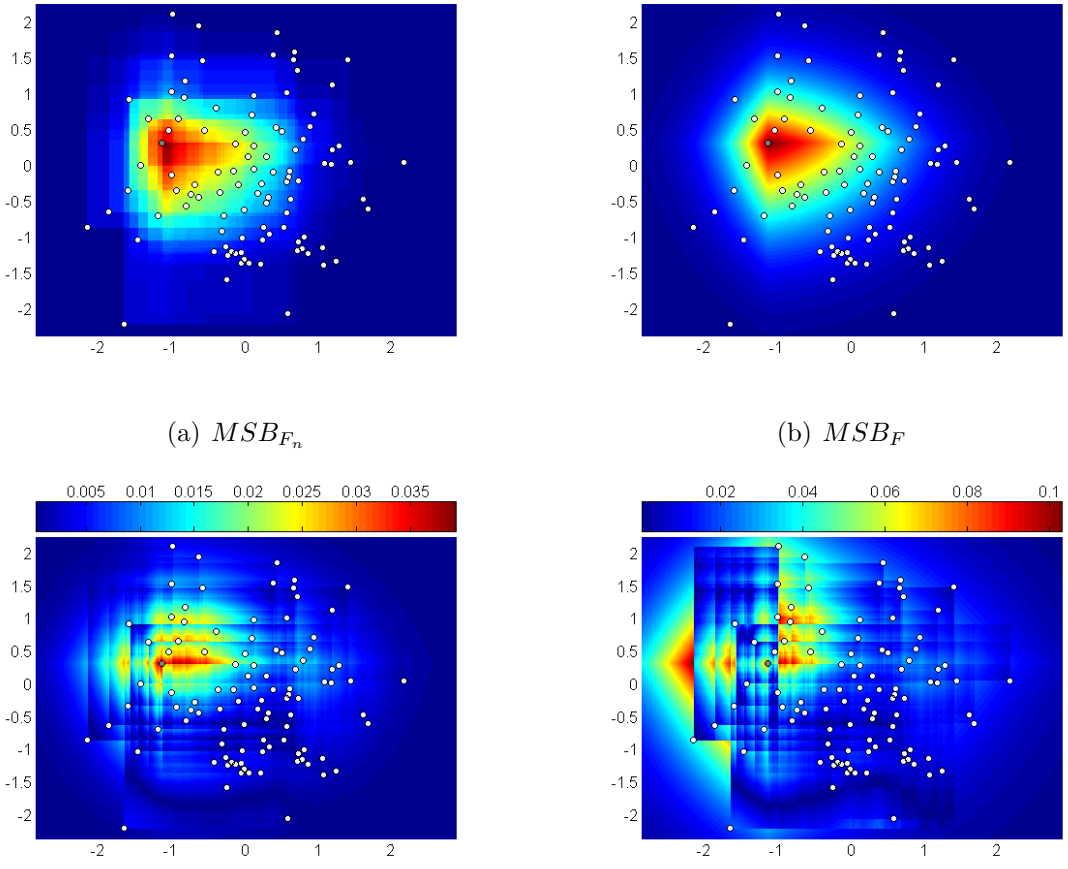
(a) MSB_{F_n}

(b) MSB_F

(c) $|MSB_F - MSB_{F_n}|$

(d) $|MSBE_F - MSBE_{F_n}|$

Figura 4.20: *Muestra de distribución normal y punto fijo central.*



(a) MSB_{F_n} (b) MSB_F

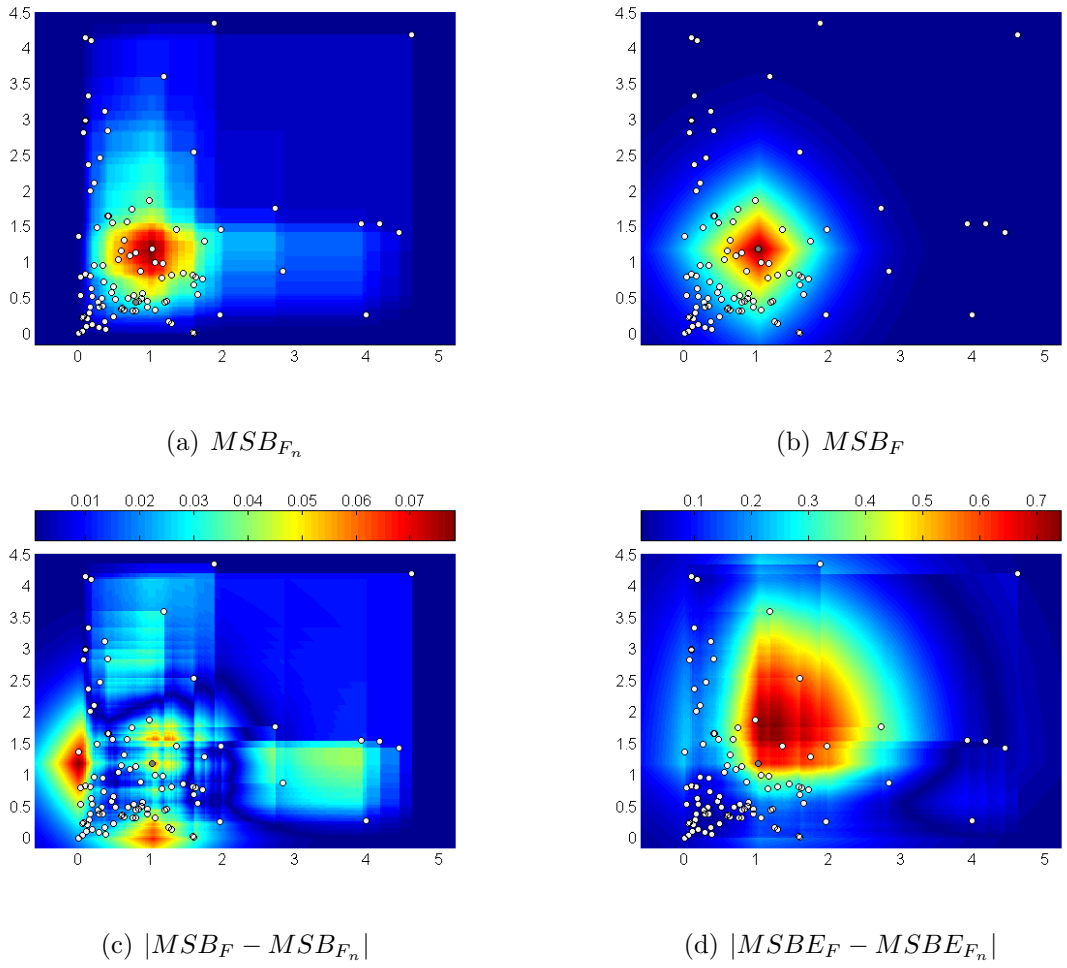
(c) $|MSB_F - MSB_{F_n}|$ (d) $|MSBE_F - MSBE_{F_n}|$

Figura 4.21: *Muestra de distribución normal y punto fijo externo.*

mayor simetría en torno al punto fijo. Estas diferencias se plasman en las superficies de $|MSB_F - MSB_{F_n}|$ y $|MSBE_F - MSBE_{F_n}|$ donde se aprecian zonas amplias con valores de la diferencia elevados, llegando en el caso del punto fijo (externo) a valores de 0.08 y 0.7 (0.1 y 0.2), respectivamente.

Las Figuras 4.24 y 4.25 corresponden a la muestra con coordenadas independientres de distribución uniforme. Las diferencias entre las similaridades muestrales y las teóricas son muy parecidas al caso de la muestra normal tanto para el punto central (4.24(a) y 4.24(b)) como para el externo (4.25(a) y 4.25(b)). La mayor diferencia con el caso normal que se aprecia es la velocidad con que decrece la similaridad muestral, ya que en el caso uniforme, al rellenarse el cuadrado de forma homogénea, se tiene que para puntos alejados del punto fijo la similaridad es mayor que en el caso normal donde resulta menos probable encontrar puntos a esa distancia. Esto implica que los valores de $|MSB_F - MSB_{F_n}|$ y $|MSBE_F - MSBE_{F_n}|$ no sean mucho más elevados que los que se obtuvieron en los ejemplos de la muestra normal (0.065 y 0.055 frente a 0.04 para $|MSB_F - MSB_{F_n}|$ y 0.12 y 0.18 frente a 0.1 para $|MSBE_F - MSBE_{F_n}|$). Sin embargo, si se observa que las zonas donde los valores son elevados tienen un área mayor que en el caso normal.

Las superficies para la muestra de distribución normal contaminada con otra normal de media $(2.5, 2.5)'$ y varianza 0.25 se encuentran en las Figuras 4.26 y 4.27. Se observa cómo los contornos para el punto fijo central están orientados hacia el grupo de contaminación y que para el externo están orientados hacia el grupo mayoritario. Las similaridades muestrales son más sensibles que las teóricas, ya que pueden encontrarse puntos de otros grupos con valores de similaridad elevados. En cuanto a la diferencia $|MSB_F - MSB_{F_n}|$ para el punto fijo central (Figura 4.26(c)) se aprecia que el máximo está muy próximo al obtenido para la muestra normal y que el área de las zonas con estos valores es elevada. No ocurre lo mismo para $|MSBE_F - MSBE_{F_n}|$ (Figura 4.26(d)) donde el máximo es 0.2 y las zonas en las que se alcanza dicho valor son pequeñas. Para el punto externo los máximos y las zonas con valores altos son mayores que en el caso normal, ya que los puntos externos tienen una profundidad teórica pequeña y al estandarizar hacen que algunas filas y columnas de la matriz presenten valores elevados.



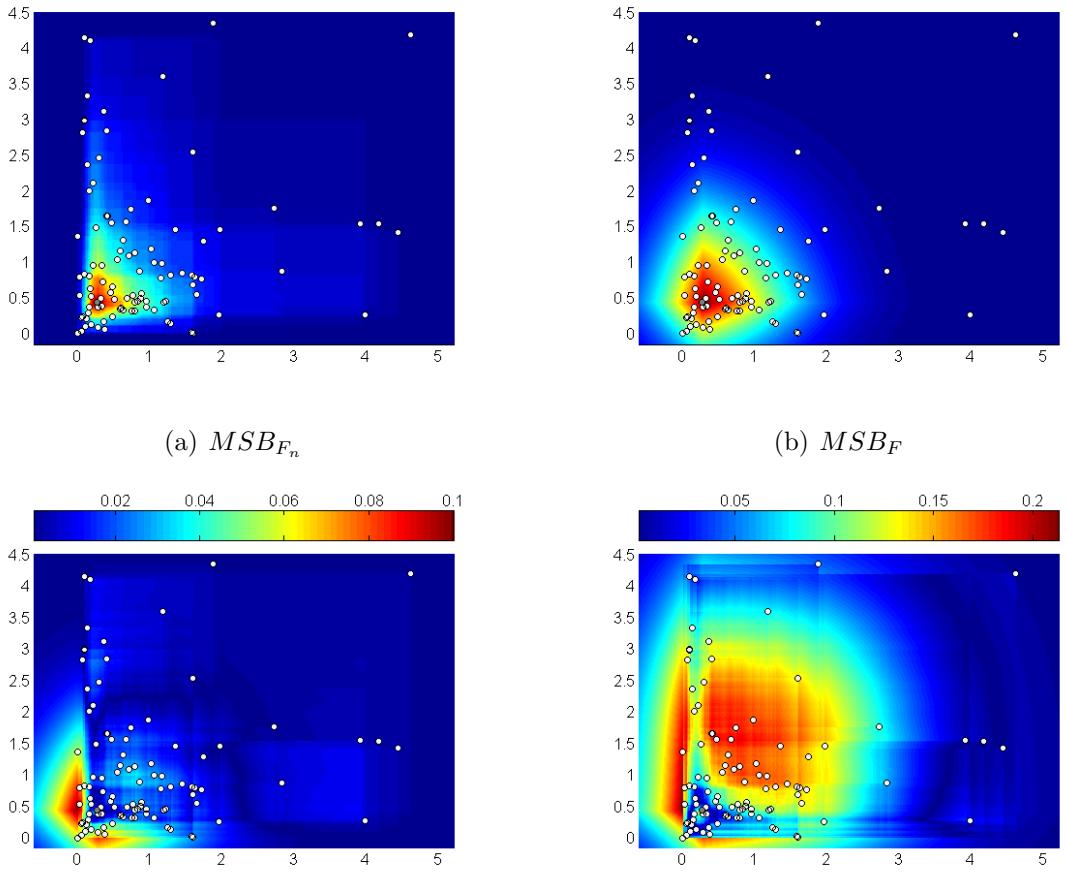
(a) MSB_{F_n}

(b) MSB_F

(c) $|MSB_F - MSB_{F_n}|$

(d) $|MSBE_F - MSBE_{F_n}|$

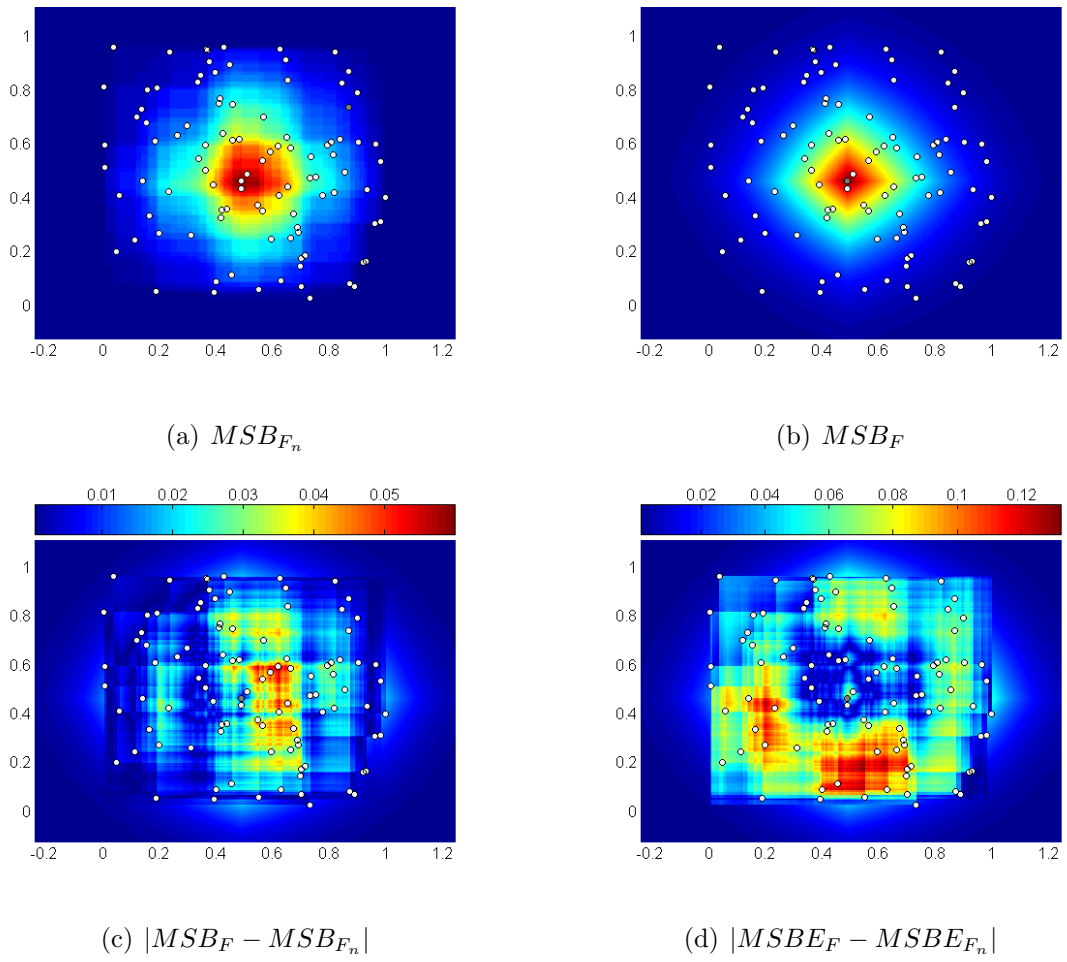
Figura 4.22: *Muestra de distribución exponencial y punto fijo central.*



(a) MSB_{F_n} (b) MSB_F

(c) $|MSB_F - MSB_{F_n}|$ (d) $|MSBE_F - MSBE_{F_n}|$

Figura 4.23: *Muestra de distribución exponencial y punto fijo externo.*



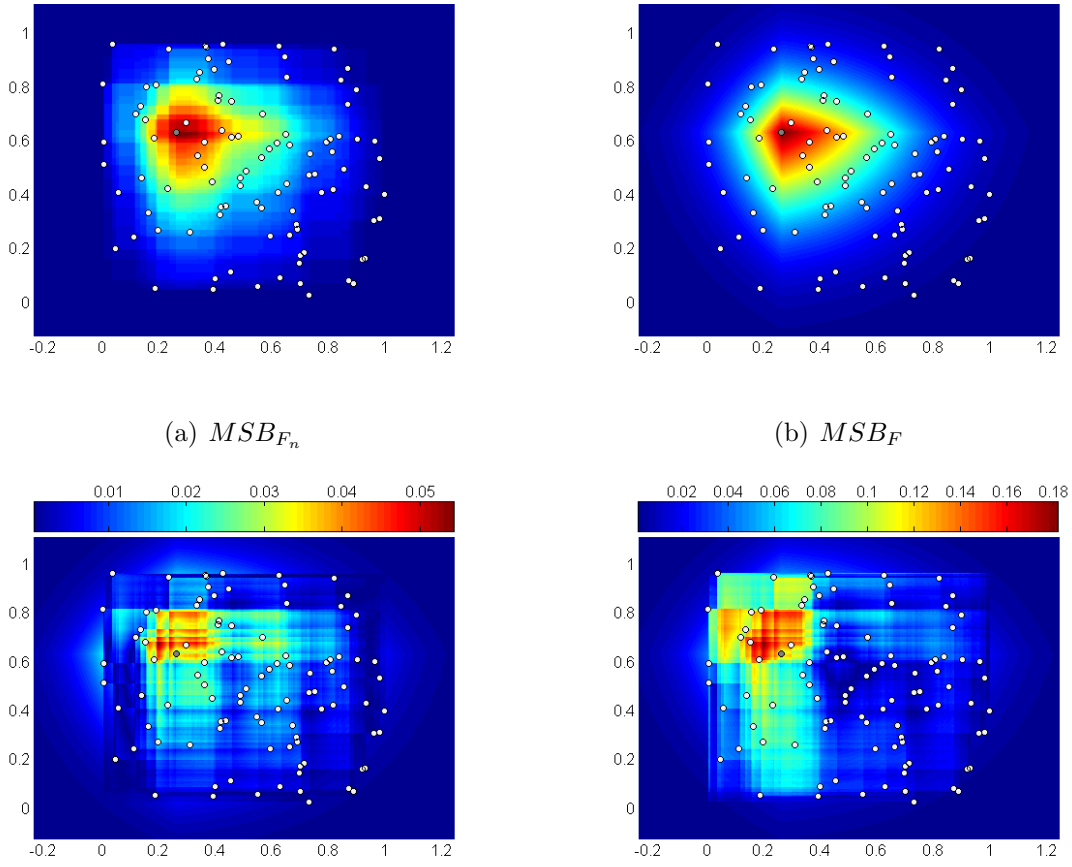
(a) MSB_{F_n}

(b) MSB_F

(c) $|MSB_F - MSB_{F_n}|$

(d) $|MSBE_F - MSBE_{F_n}|$

Figura 4.24: *Muestra de distribución uniforme y punto fijo central.*



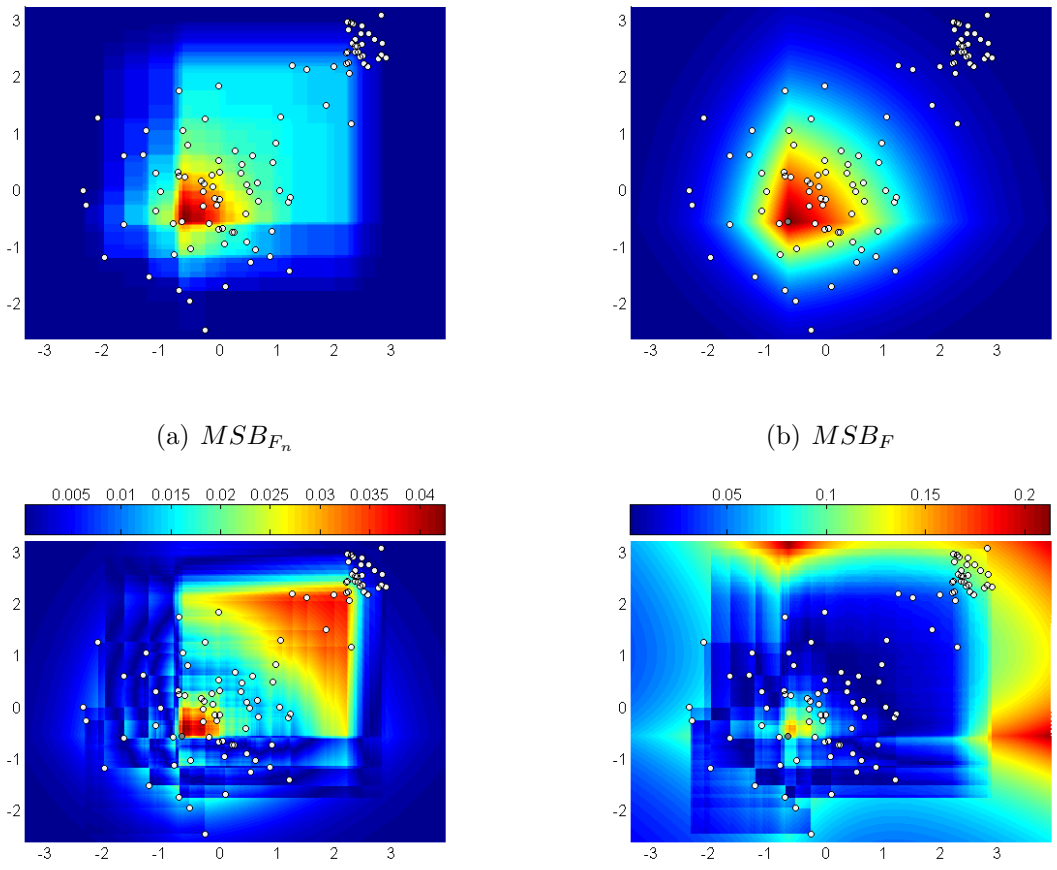
(a) MSB_{F_n}

(b) MSB_F

(c) $|MSB_F - MSB_{F_n}|$

(d) $|MSBE_F - MSBE_{F_n}|$

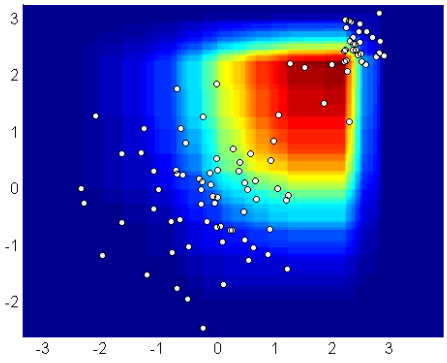
Figura 4.25: *Muestra de distribución uniforme y punto fijo externo.*



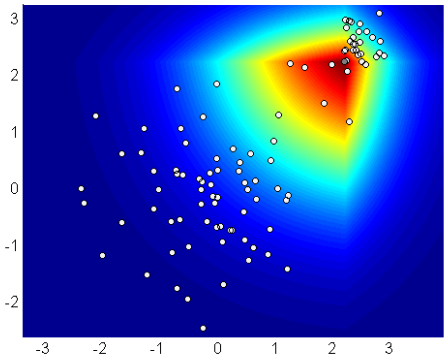
(a) MSB_{F_n} (b) MSB_F

(c) $|MSB_F - MSB_{F_n}|$ (d) $|MSBE_F - MSBE_{F_n}|$

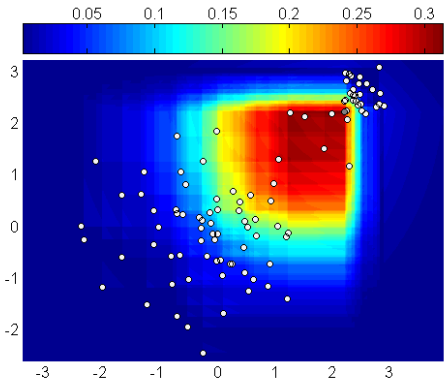
Figura 4.26: *Muestra de distribución normal contaminada y punto fijo central.*



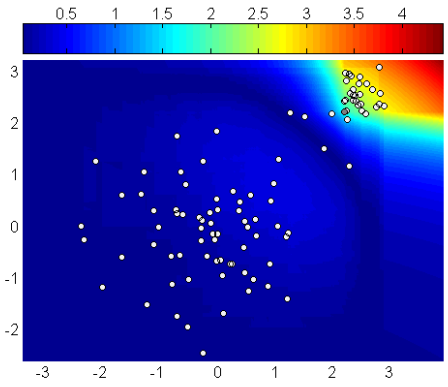
(a) MSB_{F_n}



(b) MSB_F



(c) $|MSB_F - MSB_{F_n}|$



(d) $|MSBE_F - MSBE_{F_n}|$

Figura 4.27: *Muestra de distribución normal contaminada y punto fijo externo.*

4.4.2. Valores críticos

Los valores críticos de ambos estadísticos del contraste se han obtenido mediante la simulación de 5000 muestras estandarizadas de la distribución normal estándar bivariante. Salvo para la profundidad de Oja, para la que la similaridad teórica ha sido estimada mediante muestras bivariantes estandarizadas de tamaño 10000, las similaridades teóricas son exactas. Las Tabla 4.37 contiene, para los estadísticos DS y DSE, los valores críticos para el contraste con significación 0.1, 0.05 y 0.01 y muestras de tamaño 50 y 100.

		DS			DSE		
		Percentil			Percentil		
Similaridad	Tamaño muestral	0.90	0.95	0.99	0.90	0.95	0.99
Oja	50	0.0110	0.0120	0.0160	0.0070	0.0080	0.0090
	100	0.0070	0.0080	0.0100	0.0040	0.0050	0.0060
Proyecciones	50	0.0360	0.0400	0.0500	0.0450	0.0520	0.0650
	100	0.0240	0.0270	0.0330	0.0300	0.0330	0.0410
Bandas	50	0.0120	0.0130	0.0150	0.0610	0.0650	0.0710
	100	0.0070	0.0080	0.0090	0.0450	0.0470	0.0520
Bandas modificada	50	0.0230	0.0240	0.0270	0.0500	0.0530	0.0590
	100	0.0160	0.0170	0.0190	0.3880	0.4100	0.4560

Tabla 4.37: *Valores críticos para los estadísticos DS y DSE para distribución nula normal.*

Los valores críticos de los dos estadísticos para las distribuciones nulas uniforme y exponencial se encuentran recogidos en las Tablas 4.38 y 4.39.

4.4.3. Potencia del contraste

Como en los contrastes anteriores, se realiza a continuación un estudio de la potencia para cada una de las tres distribuciones nulas. La potencia se ha estimado por medio de la simulación de 1000 muestras estandarizadas de cada alternativa. Se presentan los resultados para tamaños muestrales 50 y 100. De nuevo, salvo para la similaridad de Oja, para la que la similaridad teórica ha sido estimada mediante muestras de tamaño 10000, los cálculos de similaridad teórica son exactos. Se calcula además el índice para medir la efectividad de cada similaridad frente al resto. En cada tabla de esta sección se presentan

		DS			DSE		
		Percentil			Percentil		
Similaridad	Tamaño muestral	0.90	0.95	0.99	0.90	0.95	0.99
Oja	50	0.0051	0.0057	0.0069	0.0029	0.0032	0.0038
	100	0.0037	0.0041	0.0049	0.0020	0.0023	0.0027
Proyecciones	50	0.0320	0.0365	0.0458	0.0463	0.0529	0.0647
	100	0.0224	0.0254	0.0322	0.0320	0.0364	0.0446
Bandas	50	0.0182	0.0198	0.0231	0.0770	0.0814	0.0905
	100	0.0106	0.0117	0.0138	0.0533	0.0562	0.0631
Bandas modificada	50	0.0396	0.0438	0.0524	0.0708	0.0766	0.0884
	100	0.0251	0.0280	0.0334	0.0463	0.0499	0.0573

Tabla 4.38: *Valores críticos para los estadísticos DS y DSE para distribución nula uniforme.*

		DS			DSE		
		Percentil			Percentil		
Similaridad	Tamaño muestral	0.90	0.95	0.99	0.90	0.95	0.99
Oja	50	0.0271	0.0322	0.0430	0.0125	0.0144	0.0181
	100	0.0186	0.0217	0.0292	0.0083	0.0095	0.0120
Proyecciones	50	0.0311	0.0349	0.0425	0.0468	0.0530	0.0662
	100	0.0202	0.0227	0.0281	0.0301	0.0343	0.0431
Bandas	50	0.0146	0.0161	0.0193	0.0678	0.0721	0.0812
	100	0.0088	0.0097	0.0116	0.0491	0.0521	0.0585
Bandas modificada	50	0.0289	0.0318	0.0397	0.0560	0.0600	0.0688
	100	0.0197	0.0220	0.0275	0.0396	0.0423	0.0488

Tabla 4.39: *Valores críticos para los estadísticos DS y DSE para distribución nula exponencial.*

para cada similaridad los resultados para ambos estadísticos de contraste. Se comienza con la distribución normal.

4.4.3.1. Potencia para la distribución nula normal

La Tabla 4.40 contiene los resultados para las distribuciones alternativas compuestas por dos coordenadas independientes e igualmente distribuidas. En ésta se puede obser-

var que las distribuciones para las que se obtiene mejor potencia son la exponencial, la lognormal y la t con dos grados de libertad. También puede apreciarse que para algunas alternativas los resultados no son homogéneos para las distintas profundidades; así se tiene que para las distribuciones $gamma(5,1)$, χ^2 y beta(1,2), la similaridad por proyecciones se comporta notablemente peor que las demás. Para beta(1,1) y beta(2,2), la diferencia es muy clara en favor de la similaridad por bandas y por bandas modificada, mientras que para la logística es a la inversa. Por útimo se tiene que por lo general la potencia alcanzada con el estadístico DSE es superior a la obtenida con DS. Si una de las coordenadas se distribuye según una normal (Tabla 4.41) se tiene que las similaridades por bandas y por bandas modificada tienen una mayor presencia, ya que cuando la otra coordenada tiene distribución beta, tanto la similaridad de Oja como la de proyecciones presentan potencias prácticamente nulas.

Tomando como medida global de potencia sobre el grupo el índice, Tabla 4.42, se tiene por un lado que, para todas las similaridades, el estadístico DSE se comporta mejor que el DS y por otro que las mejores similaridades son la de Oja y la de bandas modificada.

Sobre mixturas de normales, Tabla 4.43, salvo para la similaridad por bandas modificada y, en menor medida, por bandas, el contraste no es capaz de detectar discrepancias cuando las distribuciones son diferentes sólo en media. Si los cambios se producen también en la forma de la matriz de covarianzas, la eficacia se invierte, siendo mejores la de Oja y la de proyecciones. De forma global sobre este grupo (Tabla 4.44) se tiene que los dos estadísticos para la similaridad de Oja son mejores que para el resto de similaridades y se confirma para cada similaridad que el estadístico DSE es más potente.

Para el grupo de alternativas esféricamente simétricas (Tabla 4.45) se tiene un mejor comportamiento para las distribuciones de Pearson de tipo VII y para las esféricas de radio exponencial y beta(1,2). De forma global (Tabla 4.46), se tiene que la mejor similaridad es la de bandas modificada, que para las alternativas en las que el resto no detectan nada está claramente por encima. Cuando el tamaño muestral muestral es igual a 100, se comporta mejor el estadístico DS.

Por último, para las distribuciones con correlación radial/angular, puede observarse

		Similaridad							
		SO		SP		SB		SBM	
Distribución	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
Exponencial	50	100	100	80	85	95	100	100	100
	100	100	100	98	99	100	100	100	100
Lognormal	50	100	100	99	100	100	100	100	100
	100	100	100	100	100	100	92	100	100
Gamma(5,1)	50	47	55	15	16	10	44	28	45
	100	80	90	23	26	45	80	66	82
chi-cuadrado(5)	50	79	87	30	34	34	79	66	82
	100	99	100	50	55	91	99	98	100
chi-cuadrado(15)	50	31	38	9	10	9	31	15	25
	100	62	76	16	18	26	60	42	60
t(2)	50	99	99	98	97	90	98	96	96
	100	100	100	100	100	100	92	100	100
t(5)	50	60	55	48	45	14	46	33	37
	100	84	80	71	71	52	70	67	64
Logística(0,1)	50	35	32	26	23	4	24	11	16
	100	54	48	44	43	18	40	31	30
Beta(1,1)	50	0	1	0	0	73	31	99	83
	100	2	20	0	0	99	91	100	100
Beta(1,2)	50	15	37	1	2	56	56	89	85
	100	64	94	0	1	98	96	100	100
Beta(2,2)	50	0	0	0	0	33	4	56	18
	100	0	1	0	0	68	20	89	56

Tabla 4.40: *Potencia del contraste basado en similaridades con distribución nula normal, para vectores bidimensionales cuyas componentes son independientes e igualmente distribuidas.*

		Similaridad							
		SO		SP		SB		SBM	
Distribución	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
Normal(0,1) y Exponencial	50	86	91	39	46	41	83	84	92
	100	100	100	68	76	96	99	100	100
Normal(0,1) y chi-cuadrado(5)	50	47	54	14	15	11	44	30	46
	100	79	87	22	27	47	79	70	84
Normal(0,1) y t(5)	50	32	30	21	20	5	22	11	19
	100	53	50	34	36	15	34	29	33
Normal(0,1) y Beta(1,1)	50	1	3	0	1	26	14	58	35
	100	1	8	0	1	67	50	93	84
Normal(0,1) y Beta(1,2)	50	9	16	2	3	23	27	45	42
	100	21	47	1	2	62	62	86	87

Tabla 4.41: *Potencia del contraste basado en similaridades con distribución nula normal, para vectores bidimensionales cuyas componentes son independientes y poseen distribuciones diferentes.*

	Profundidad de ordenación							
	SO		SP		SB		SBM	
n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
50	3.53	2.94	6.53	5.97	5.91	3.69	3.97	3.47
100	3.81	3.09	6.59	5.97	5.00	4.53	3.75	3.25

Tabla 4.42: *Índice de rangos para el contraste basado en similaridades con distribución nula normal y el grupo 1 de alternativas.*

		Similaridad							
		SO		SP		SB		SBM	
Distribución	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
Mixtura Normal (2,0,0)	50	1	3	1	1	3	4	12	7
	100	1	5	0	1	4	8	17	11
Mixtura Normal (4,0,0)	50	4	12	1	3	22	40	90	83
	100	9	44	0	5	59	86	100	100
Mixtura Normal (2,0.9,0)	50	77	84	43	45	21	35	24	44
	100	98	100	78	82	56	67	64	80
Mixtura Normal (0.5,0.9,0)	50	52	52	44	44	21	17	6	12
	100	81	82	73	76	34	22	14	19
Mixtura Normal (0.5,0.9,-0.9)	50	98	99	99	98	93	63	9	9
	100	100	100	100	100	100	97	12	12

Tabla 4.43: *Potencia del contraste basado en similaridades con distribución nula normal, para mixturas bidimensionales de normales.*

	Profundidad de ordenación							
	SO		SP		SB		SBM	
n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
50	4.00	2.70	5.00	4.80	5.30	4.80	4.90	4.50
100	3.90	2.80	5.60	4.50	5.00	4.80	5.00	4.40

Tabla 4.44: *Índice de rangos para el contraste basado en similaridades con distribución nula normal y el grupo 2 de alternativas.*

		Similaridad							
		SO		SP		SB		SBM	
Distribución	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
PearsonII(0)	50	0	0	0	0	3	4	84	36
	100	24	40	0	3	14	65	99	85
PearsonII(1)	50	0	0	0	0	6	1	42	13
	100	1	3	0	0	11	6	72	30
PearsonVII(2)	50	99	99	98	98	86	95	94	95
	100	100	100	100	100	99	94	100	100
PearsonVII(3)	50	83	80	69	66	27	57	46	51
	100	97	95	93	93	68	80	83	81
PearsonVII(5)	50	42	38	29	27	6	24	11	15
	100	66	59	51	52	17	33	29	29
Esférica(Exponencial)	50	100	100	100	100	92	98	98	99
	100	100	100	100	100	100	100	100	100
Esférica(Gamma(5,1))	50	1	3	0	0	6	5	20	10
	100	1	4	0	0	7	10	34	22
Esférica(Beta(1,1))	50	8	7	30	31	6	6	14	23
	100	9	10	42	45	9	5	35	45
Esférica(Beta(1,2))	50	75	72	83	84	30	49	48	62
	100	96	96	98	98	73	69	88	91
Esférica(Beta(2,2))	50	0	0	1	1	6	1	22	7
	100	0	0	0	0	10	3	45	16

Tabla 4.45: *Potencia del contraste basado en similaridades con distribución nula normal, para distribuciones de Pearson y esféricas.*

	Profundidad de ordenación							
	SO		SP		SB		SBM	
n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
50	4.05	4.35	4.15	4.15	6.05	5.15	4.25	3.85
100	4.35	4.10	4.95	4.60	5.70	5.45	3.35	3.50

Tabla 4.46: *Índice de rangos para el contraste basado en similaridades con distribución nula normal y el grupo 3 de alternativas.*

en la Tabla 4.47, cómo las similaridades se comportan de forma más homogénea que en los grupos anteriores y cómo la potencia aumenta conforme el coeficiente de correlación se acerca a 1, caso en el que la mayoría de las similaridades obtiene un rechazo cercano al 100 %. Las mejores similaridades en este grupo (Tabla 4.48) son, con mucha diferencia sobre el resto, la de Oja y la de bandas modificada en su versión DSE.

		Similaridad							
		SO		SP		SB		SBM	
Distribución	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
0.2	50	7	8	5	6	6	9	6	9
	100	10	16	6	7	7	11	12	16
0.4	50	17	24	11	14	5	18	14	23
	100	37	59	12	17	18	36	31	46
0.6	50	38	58	15	25	10	40	33	49
	100	79	95	21	37	46	73	74	87
0.8	50	66	85	23	40	20	66	61	79
	100	98	100	37	65	80	96	98	100
1	50	89	98	38	65	41	92	89	96
	100	100	100	63	90	98	100	100	100

Tabla 4.47: *Potencia del contraste basado en similaridades con distribución nula normal, para distribuciones con correlación radial/angular.*

	Profundidad de ordenación							
	SO		SP		SB		SBM	
n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
50	4.00	1.40	7.40	5.90	7.40	2.80	5.20	1.90
100	3.50	1.60	8.00	6.90	6.10	4.20	3.70	2.00

Tabla 4.48: *Índice de rangos para el contraste basado en similaridades con distribución nula normal y el grupo 4 de alternativas.*

De forma global para las 36 alternativas (Tabla 4.49) se tiene por un lado que la similaridad que mejor se comporta es la de Oja, seguida de cerca por la de bandas modificada y por otro, que se tiene una mayor potencia si se emplea la diferencia entre

matrices de similaridades estandarizadas.

	Profundidad de ordenación							
	SO		SP		SB		SBM	
n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
50	3.81	3.08	5.78	5.29	6.07	4.13	4.35	3.50
100	3.93	3.13	6.19	5.51	5.35	4.78	3.81	3.31

Tabla 4.49: *Índice de rangos global para el contraste basado en similaridades con distribución nula normal.*

4.4.3.2. Potencia para la distribución nula uniforme

Para vectores aleatorios de coordenadas independientes y con distribución beta, se tiene que tanto la similaridad por bandas, como la similaridad por bandas modificada, presentan porcentajes de rechazo más elevados para parámetros menores que uno. Mientras que, para valores mayores, la de Oja y la de proyecciones se comporta mejor. Para la uniforme en la circunferencia unidad, la similaridad por bandas para el estadístico DSE obtiene los mejores resultados. Tras ésta, se sitúa la de Oja, con el estadístico DS. Para la normal estándar sólo la similaridad por proyecciones presenta rechazos en la práctica totalidad de las muestras para ambos estadísticos. En el extremo opuesto se encuentra la similaridad por bandas con el que apenas rechaza un 85 % de las ocasiones. Sobre distribuciones normales truncadas fuera de circunferencias, tanto la similaridad por bandas como la de bandas modificada se muestran muy superiores al resto.

Tanto para las mixturas de distribuciones uniformes, como para las distribuciones uniformes en cuadrados recortados (Tabla 4.51), son las similaridades por proyecciones y de Oja las que obtienen porcentajes de rechazo mayores. Para las distribuciones de Pearson tipo II, de nuevo la similaridad de Oja aparece como una de las mejores (DS), acompañada por la de bandas (DSE) que obtiene porcentajes mayores que ésta. Para este conjunto de alternativas, con el estadístico DS se obtienen siempre mejores resultados.

Globalmente, sobre las 20 distribuciones alternativas, se obtienen resultados heterogéneos (véase la Tabla 4.52): existe una elevada diferencia entre la potencia media

		Similaridad							
		SO		SP		SB		SBM	
Distribución	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
Beta(0.8,0.8)	50	18	11	2	3	29	26	40	39
	100	39	17	1	3	53	46	66	69
Beta(0.9,0.9)	50	8	6	2	3	13	10	16	16
	100	9	6	1	3	20	14	25	26
Beta(1.15,1.15)	50	5	4	12	10	2	4	1	1
	100	13	6	18	12	1	9	2	3
Beta(1.3,1.3)	50	14	5	22	17	0	9	0	1
	100	35	9	38	28	5	30	11	15
Unif. Circunferencia	50	17	6	14	9	0	32	1	2
	100	59	18	27	16	46	98	36	39
Normal	50	99	93	99	99	85	99	92	96
	100	100	100	100	100	100	100	100	100
Normal circ. (1)	50	38	9	31	20	3	51	4	7
	100	87	43	68	50	80	99	77	78
Normal circ. (1.5)	50	66	22	60	45	11	67	17	24
	100	98	77	93	86	96	100	98	98
Normal circ. (2)	50	87	46	87	76	35	85	52	59
	100	100	98	100	99	100	100	100	100

Tabla 4.50: *Potencia del contraste basado en similaridades con distribución nula uniforme y alternativas beta, uniforme en circunferencia y normal.*

		Similaridad							
		SO		SP		SB		SBM	
Distribución	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
Mixt. Unif. (0.1,0.5)	50	10	12	7	8	6	6	5	4
	100	15	15	10	12	5	5	6	6
Mixt. Unif. (0.25,0.5)	50	45	48	21	25	6	9	5	4
	100	77	81	46	42	11	11	7	6
Mixt. Unif. (0.1,0.25)	50	20	24	9	15	25	16	20	14
	100	42	44	20	25	39	23	35	25
Mixt. Unif. (0.25,0.25)	50	87	89	65	67	69	41	65	31
	100	99	100	95	94	96	67	94	63
Cuad. Recort (0.2)	50	5	5	6	6	2	3	3	3
	100	7	7	7	6	1	4	2	2
Cuad. Recort (0.4)	50	14	13	10	8	4	7	1	1
	100	26	23	13	12	2	16	3	4
Cuad. Recort (0.6)	50	54	51	22	20	11	15	1	1
	100	93	93	46	43	15	23	8	7
Cuad. Recort (0.8)	50	97	97	51	56	34	27	4	2
	100	100	100	93	91	60	35	28	16
Cuad. Recort (1)	50	100	100	71	83	65	45	12	6
	100	100	100	100	100	94	51	71	39
Pearson II (0)	50	18	5	15	9	0	30	1	2
	100	63	18	27	16	49	98	37	41
Pearson II (1)	50	75	30	69	56	19	77	29	35
	100	100	93	98	95	99	100	99	99

Tabla 4.51: *Potencia del contraste basado en similaridades con distribución nula uniforme y alternativas mixtura de uniforme, uniforme en cuadrados recortados y Pearson II.*

de la mejor similaridad (Oja) con respecto a la peor (bandas modificada) y, para las similaridades de Oja y por bandas, los resultados dependen del estadístico elegido.

		Profundidad de ordenación							
		SO		SP		SB		SBM	
	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
Potencia media	50	43.85	33.8	33.75	31.75	20.95	32.95	18.45	17.40
	100	63.1	52.40	50.05	46.65	48.60	51.45	45.25	41.80
Global	50	2.40	3.85	3.70	3.80	6.05	3.88	6.18	6.15
	100	2.42	4.43	4.47	5.35	4.95	4.08	5.13	5.18

Tabla 4.52: *Porcentaje medio de rechazo e índice de rangos para el contraste basado en similaridades sobre distribución nula uniforme.*

4.4.3.3. Potencia para la distribución nula exponencial

A continuación se introducen los resultados obtenidos para la distribución nula exponencial. En la Tabla 4.53, se encuentran las distribuciones alternativas obtenidas a partir de la normal, la chi-cuadrado y la gamma. Para la distribución normal se tiene que todas las similaridades, salvo la de Oja, rechazan la totalidad de las muestras. El comportamiento opuesto se da para su valor absoluto, ya que sólo esta similaridad discrimina correctamente éste y la distribución exponencial. Las similaridades por proyecciones y por bandas obtienen los porcentajes de rechazo más elevados para la lognormal. Para los vectores aleatorios con componentes distribuidos según la chi-cuadrado con diferentes grados de libertad, son las similaridades de Oja y la de bandas modificada las que tienen un mejor comportamiento. Por último, para la distribución gamma de parámetros 5 y 1, todas las similaridades rechazan el 100 % de las muestras.

Para los vectores cuyas componentes tienen distribución Weibull (Tabla 4.54), si el parámetro de forma es menor que uno, las similaridades por bandas y por bandas modificada obtienen los mejores resultados. Si dicho parámetro es mayor que uno, la de Oja tiene un mejor funcionamiento. Por último, sobre las mixturas de exponenciales, de nuevo la similaridad por bandas y por bandas modificada obtienen porcentajes de rechazo mayores.

		Similaridad							
		SO		SP		SB		SBM	
Distribución	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
Normal	50	58	86	100	100	100	100	100	100
	100	49	85	100	100	100	100	100	100
|Normal|	50	90	88	11	6	7	54	30	44
	100	99	99	14	7	44	84	70	82
Lognormal	50	8	5	71	70	16	70	37	57
	100	11	8	94	93	70	96	88	97
Chi-cuadrado (1)	50	93	82	8	27	99	94	100	100
	100	100	100	34	46	100	100	100	100
Chi-cuadrado (3)	50	81	77	31	22	22	70	59	64
	100	97	96	50	34	80	94	95	95
Chi-cuadrado (4)	50	100	100	76	62	82	99	99	100
	100	100	100	94	87	100	100	100	100
Chi-cuadrado (5)	50	100	100	95	88	99	100	100	100
	100	100	100	100	99	100	100	100	100
Chi-cuadrado (10)	50	100	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100	100
Gamma (5,1)	50	100	100	100	100	100	100	100	100
	100	100	100	100	100	100	100	100	100

Tabla 4.53: *Potencia del contraste basado en similaridades con distribución nula exponencial y alternativas normal, lognormal, gamma y chi-cuadrado.*

		Similaridad							
		SO		SP		SB		SBM	
Distribución	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
Weibull (1,0.5)	50	100	100	86	96	100	100	100	100
	100	100	100	100	100	100	100	100	100
Weibull (1,0.75)	50	66	43	6	20	81	47	90	88
	100	98	95	20	31	98	83	99	99
Weibull (1,0.9)	50	5	2	4	6	21	7	27	23
	100	22	12	4	8	35	13	40	37
Weibull (1,1.1)	50	24	21	7	6	2	16	7	9
	100	43	39	11	7	8	24	18	20
Weibull (1,1.3)	50	91	88	27	19	19	69	58	65
	100	100	99	43	29	75	93	94	96
Weibull (1,1.7)	50	100	100	86	80	97	100	100	100
	100	100	100	99	97	100	100	100	100
Mixt. Expo. (0.3,0.2)	50	19	12	16	40	70	42	67	56
	100	60	67	55	70	93	76	90	87
Mixt. Expo. (0.2,0.2)	50	9	5	8	19	49	24	47	36
	100	31	29	25	35	77	45	71	62
Mixt. Expo. (0.1,0.2)	50	4	3	5	8	22	10	25	19
	100	9	6	9	14	40	16	30	25
Mixt. Expo. (0.3,0.5)	50	3	3	6	9	9	6	9	8
	100	5	4	10	13	15	7	13	12
Mixt. Expo. (0.2,0.5)	50	4	3	5	7	8	4	10	8
	100	4	3	7	8	13	5	8	7
Mixt. Expo. (0.1,0.5)	50	4	4	6	7	6	5	7	5
	100	4	4	7	7	7	5	7	6

Tabla 4.54: *Potencia del contraste basado en similaridades con distribución nula exponencial y alternativas Weibull y mixtura de exponenciales.*

De forma global sobre las 21 alternativas (Tabla 4.55), la similaridad que obtiene tanto un mayor porcentaje de rechazo medio, como un índice de rangos menor es la de bandas modificada. El estadístico DS presenta un menor índice de rangos, mientras que el DSE tiene un porcentaje medio más elevado, aunque apenas exsita diferencia entre ambas medidas.

		Profundidad de ordenación							
		SO		SP		SB		SBM	
	n	DS	DSE	DS	DSE	DS	DSE	DS	DSE
Potencia media	50	55.19	53.43	40.67	42.48	52.81	57.95	60.57	61.05
	100	63.43	64.10	51.24	51.67	69.29	68.62	72.52	72.62
Global	50	4.52	5.26	5.86	5.60	4.33	3.98	3.10	3.36
	100	4.57	5.05	5.88	5.74	3.64	4.31	3.33	3.48

Tabla 4.55: *Porcentaje medio de rechazo e índice de rangos para el contraste basado en similaridades sobre distribución nula exponencial.*

4.5. Comparación de los contrastes basados en profundidad

En esta sección se realiza la comparación de los tres contrastes que se han propuesto. La comparación para la distribución nula normal se lleva a cabo en dos direcciones. La primera persigue encontrar los casos específicos (contraste y profundidad/similaridad) de entre todas las posibilidades, que mejor comportamiento global presenten, teniendo en cuenta tanto el índice empleado en el estudio de potencia de cada contraste como la potencia media de cada posibilidad. En segundo lugar se realiza una comparación entre contrastes para cada una de las profundidades/similaridades para concluir qué contraste es más adecuado u obtiene una mayor información sobre cada una de estas funciones. Para las distribuciones nulas uniforme y exponencial, debido a que no se ha aplicado el contraste de la curva de concordancia para todas las funciones de profundidad, sólo se realizará la primera comparación de las comentadas.

4.5.1. Comparación global

La primera comparación que se lleva a cabo es la global, cuyo objetivo es encontrar las combinaciones de profundidad-similaridad y tipo de contraste para las que se obtienen mejores resultados de potencia. Para la determinación de un ranking de combinaciones se tiene en cuenta tanto el índice de rangos como la probabilidad de rechazo media sobre las 36 alternativas. Se comienza con un análisis para los contrastes aplicados sobre la distribución nula normal.

Los resultados para el índice y la probabilidad media se encuentran, respectivamente, en las Tablas 4.56 y 4.57. En éstas se dispone de las medidas para cada grupo de distribuciones y para tamaños muestrales 50 y 100. Adicionalmente, en ambas tablas, con el fin de ayudar a interpretar los resultados, se han resaltado con tres tonos de gris las nueve mejores combinaciones de cada fila de la tabla. Cuanto mejor es la combinación, más oscuro es el color de la celda.

En la Tabla 4.56, correspondiente al índice de rangos, puede observarse cómo el contraste de la curva de escala tiene un comportamiento peor que las otras dos opciones. Apenas tiene celdas coloreadas y las que tiene generalmente poseen tonos gris claro. Entre los otros dos contrastes se aprecia una leve diferencia en favor de los basados en similaridad, para los que la mayoría de las celdas sombreadas lo están en color gris oscuro. Por grupos de distribución se tiene que, para el primero, los contrastes basados en similaridad son considerablemente mejores. Para el segundo grupo de distribuciones, el de mixturas de normales, el mejor es el contraste basado en las curvas de concordancia, seguido de cerca por el basado en similaridades. Para el grupo de alternativas esféricamente simétricas (grupo tercero), las similaridades no son capaces de detectar las discrepancias entre la muestra y la distribución normal, mientras que el contraste basado en concordancia sí lo es. El de la curva de escala se sitúa lejos de éste pero por encima del anterior. Para el último grupo, el de distribuciones con correlación radial/angular, de nuevo los basados en similaridades se muestran superiores al resto. En cuanto a las profundidades, se tiene que la profundidad/similaridad que mejor comportamiento ofrece es la de Oja, ya que posee celdas sombreadas para todos los contrastes.

Para el basado en la curva de concordancia, además de ésta, también aparecen la profundidad semiespacial, la de proyecciones y, en menor medida, la L_1. Por último, en el contraste basado en similaridades destacan tanto la de Oja como la de bandas modificada. Sobre las 36 alternativas, para tamaño muestral 50, la mejor combinación es DSE/Oja, seguida de ACC/Oja y DS/Oja y para tamaño muestral 100, en primer lugar está DSE/Oja, seguida de DS/Oja, ACC/Oja y ACC/Proyecciones.

Para la probabilidad de rechazo media, Tabla 4.57, se pueden apreciar algunos cambios relevantes como que el contraste de la curva de escala ofrece mejores resultados que para el índice en detrimento del contraste basado en la curva de concordancia y del basado en la similaridad sin estandarizar. Se observa como el contraste basado en similaridades sigue siendo mejor que el resto para los grupos 1 y 4, y que el basado en concordancia sobresale solamente para el grupo 2 de distribuciones, ya que sobre el grupo 3 el que mejor se comporta en términos de potencia es el de la curva de escala. De forma global sobre todas las alternativas destaca el de similaridades, ya que para muestras de tamaño 50 obtiene los tres valores más altos y cuando el tamaño es 100 obtiene 2 de los tres valores más elevados. En cuanto a profundidades/similaridades la que obtiene medias más altas es la de bandas modificada. La profundidad y similaridad de Oja no obtiene resultados tan positivos como anteriormente si bien sigue destacando de forma global en todos los contrastes.

A continuación se analizan los tres contrastes sobre las tres distribuciones nulas para las que se ha aplicado el contraste. Se comparan, para cada tamaño muestral, tanto el índice de rangos, como el porcentaje de rechazo globales sobre todas las alternativas de cada distribución. La Tabla 4.58 contiene estas medidas para cada combinación contraste / profundidad (similaridad).

Se observa cómo ambas medidas de bondad de los contrastes ofrecen ordenaciones similares, es decir, las celdas sombreadas para el índice de rangos son, en su mayoría, las mismas que para el porcentaje de rechazo, si bien si se producen, para algunas combinaciones cambios sustanciales en la posición dentro de cada fila. Esto se debe a que hay combinaciones que se comportan razonablemente bien en comparación al resto, para cada

		A($C_{n,p}$)							ACC							DS				DSE			
	n	PSem	PS	PO	PP	PL_1	PB	PBM	PSem	PS	PO	PP	PL_1	PB	PBM	SO	SP	SB	SBM	SO	SP	SB	SBM
Grupo 1	50	15.13	14.13	10.41	11.41	11.50	13.41	15.47	10.00	14.53	8.50	10.06	10.69	11.22	11.16	7.44	16.94	13.84	7.78	6.44	16.28	8.59	8.09
	100	13.41	13.38	9.72	10.50	11.66	12.75	15.22	11.06	14.88	10.06	10.56	11.75	10.31	12.47	8.47	17.41	9.78	7.66	7.25	16.03	10.91	7.78
Grupo 2	50	16.30	14.60	11.00	11.90	11.80	13.10	15.30	5.70	13.30	7.20	8.60	9.90	12.30	16.50	8.50	9.80	14.10	12.30	7.10	9.30	12.40	12.00
	100	15.50	14.50	10.50	12.30	12.00	11.70	15.60	6.50	12.90	8.90	8.80	10.80	11.60	14.90	9.20	11.10	12.30	12.50	7.40	9.80	12.00	12.20
Grupo 3	50	10.50	10.30	7.70	8.35	7.25	9.55	13.50	6.75	11.45	7.00	6.95	8.20	11.20	12.45	12.45	14.55	19.00	14.50	13.55	14.45	18.00	15.35
	100	9.05	8.45	7.95	9.55	8.30	7.20	13.80	7.80	12.55	6.80	7.45	9.05	10.75	11.90	13.50	15.90	18.25	13.10	14.20	15.55	17.90	14.00
Grupo 4	50	18.20	18.50	11.30	8.70	12.60	15.10	16.80	15.50	20.60	9.20	8.20	15.00	16.20	16.00	4.50	12.00	14.30	6.40	1.40	7.50	3.10	1.90
	100	15.90	14.60	9.50	9.70	14.90	11.70	19.00	14.10	19.80	13.30	9.30	18.00	19.30	19.30	3.70	12.50	7.10	3.90	1.80	9.00	4.40	2.20
Global	50	14.43	13.74	9.86	10.25	10.51	12.53	15.08	9.26	14.35	8.00	8.74	10.49	12.06	12.93	8.57	14.60	15.38	10.08	7.81	13.58	10.97	9.79
	100	12.83	12.33	9.31	10.38	11.22	10.92	15.40	9.94	14.64	9.44	9.28	11.74	11.86	13.60	9.31	15.43	12.11	9.32	8.44	14.06	12.10	9.35

Tabla 4.56: *Comparación del índice para los tres contrastes y todas las profundidades/similaridades. Distribución nula normal.*

distribución alternativa y, sin embargo, poseen porcentajes de rechazo lejos del máximo para cada una de éstas (porcentaje global intermedio e índice de rangos alto). O al revés, que tienen valores del porcentaje altos para la mayoía de las alternativas y para el resto valores muy bajos (porcentaje global alto e índice de rangos intermedio). Por ejemplo, la similaridad de Oja aplicada al estadístico DS para la distribución nula normal, posee un índice de rangos global bajo y, sin embargo, no aparece dentro de las mejores en cuanto al porcentaje de rechazo. En el lado opuesto está la similaridad por bandas también para DS, que posee un índice medio-alto y un porcentaje de rechazo de los más elevados.

Para la distribución nula normal, el contraste que mejores resultados ofrece es el basado en las similaridades de Oja y de bandas modificada, para el estadístico obtenido a partir de las matrices estandarizadas. Para la distribución uniforme, se tiene que el contraste basado en la curva de volumen es sustancialmente mejor que el resto. Su mejor resultado se obtiene ordenando según las profundidades del semiespacio, de bandas y de Oja. Le sigue el contraste basado en la similaridad de Oja con el estadístico DS. Por último, para la exponencial, el contraste basado en la similaridad por bandas y por bandas modificada, obtiene tanto los índices más bajos, como los porcentajes medios más elevados. El contraste basado en la curva de concordancia ofrece buenos resultados para contrastar normalidad para las profundidades del semiespacio, de Oja y por proyecciones. Debido a la imposibilidad de su aplicación práctica para otras distribuciones, se desconoce el desempeño fuera de esta distribución.

A la vista de los resultados, se tiene que el contraste basado en la curva de volumen es la mejor opción de todas las propuestas para variables aleatorias que tomen valor en recintos acotados, mientras que la mejor opción en caso contrario es contraste basado en las similaridades de Oja, bandas y bandas modificada.

4.5.2. Comparación individual para la distribución nula normal

El otro enfoque de comparación que se explora es el que indicará para cada función de profundidad o similaridad, cúal es el tipo de contraste con el que se obtienen mejores resultados.

La Tabla 4.59 contiene los resultados del cálculo del índice para los contrastes de cada función y los resultados del porcentaje de rechazo medio sobre las 36 alternativas. Para las profundidades en que sólo se aplican los contrastes de la curva de escala y la de concordancia, se tiene que el valor mínimo para el índice de rangos se alcanza siempre para este último. Mientras que para aquellas profundidades (similaridades) para las que se aplicaron los cuatro contrastes, los mejores resultados se obtienen para el contraste basado en la similaridad estandarizada, salvo para la profundidad por proyecciones cuyo mejor resultado se obtiene para la curva de concordancia y la de bandas modificada en muestras de 100 observaciones en que el menor índice se corresponde con el basado en la similaridad sin estandarizar.

En cuanto a los porcentajes medios de rechazo los resultados son ligeramente diferentes, quedando los cuatro contrastes más igualados. El contraste de la curva de escala es el mejor para las profundidades simplicial y L_1 y, en muestras de tamaño 100, para la de Oja y por proyecciones. El basado en la curva de escala ocupa el primer lugar para la profundidad semiespacial y, en muestras de tamaño 100, para la de proyecciones. Finalmente, el basado en la similaridad sin estandarizar es el mejor para la profundidad por bandas modificada mientras que el basado en la estandarizada lo es para la profundidad por bandas y para la de Oja sobre tamaño muestral 50.

4.6. Comparación con otros contrastes de normalidad

En esta sección se comparan las mejores combinaciones de profundidad y de tipo de contraste con algunos de los contrastes de normalidad más relevantes de la literatura. Las comparaciones se realizan teniendo en cuenta tanto el índice de rangos sobre las 36 alternativas como el porcentaje medio de rechazo. Se obtienen ambas características para cada uno de los cuatro grupos y para el total. Se toma un tamaño muestral de 50 observaciones pues es en muestras pequeñas donde los contrastes pueden presentar mayores dificultades a la hora de detectar discrepancias. Sobre muestras grandes, la mayoría

ofrecerá buenos resultados.

Las combinaciones que son objeto de comparación son los cuatro tipos de contraste para la profundidad (similaridad) de Oja, el contraste de la curva de concordancia para la profundidad por proyecciones y los contrastes basados en similaridades para la similaridad por bandas modificada.

La comparación de estos siete ejemplos se realiza frentre a los contrastes de asimetría y curtosis en Mardia (1970) (MA y MC), de Shapiro-Wilks en Fattorini (1986) (FA), de la función característica empírica en Henze y Zirkler (1990) (HZ), la segunda versión del contraste de chi-cuadrado en Quiroz y Dudley (1991) (QD), el contraste de vecinos más próximos de Zhou y Jammalamadaka (1993) (NN) y el estadístico Q en Bartoszynski et al. (1997).

La Tabla 4.60 contiene los resultados del índice de rangos y del porcentaje medio de rechazo para cada grupo de distribuciones y para el global. En cuanto al índice se tiene que el mejor contraste es el basado en la similaridad de Oja estandarizada. Tras éste están los contrastes FA, Q, el basado en la similaridad de Oja sin estandarizar. Los que peor comportamiento tienen son los basados en la asimetría y la curtosis y el de los vecinos más próximos. En cuanto al porcentaje de rechazo, se tiene que los tres primeros puestos los ocupan por orden FA, Q y HZ y detrás de estos todas las combinaciones propuestas. En este caso se tiene que los contrastes basados en la similaridad por bandas modificada se comportan mejor que los basados en la de Oja, pero no a gran distancia.

Se concluye por tanto que los contrastes basados en las matrices de similaridad tanto para Oja como para bandas modificada son competitivos en relación con alguno de los contrastes multivariantes más importantes. También que aunque el porcentaje de rechazo sea ligeramente inferior a éstos, debido a que para algún grupo de distribuciones el contraste presenta problemas, se tiene que de forma global ocupa mejores posiciones cuando se comparan entre ellos.

		$A(C_{n,p})$							ACC							DS				DSE			
	n	PSem	PS	PO	PP	PL_1	PB	PBM	PSem	PS	PO	PP	PL_1	PB	PBM	SO	SP	SB	SBM	SO	SP	SB	SBM
Grupo 1	50	32.8	33.8	45.8	44.7	44.1	39.0	34.8	45.0	31.4	45.2	44.0	42.1	35.8	35.8	46.3	30.2	39.0	57.6	49.9	31.1	50.1	57.6
	100	52.7	53.0	67.2	66.2	62.9	57.9	51.3	62.6	42.4	62.4	61.5	59.7	52.4	48.1	62.4	39.2	67.6	79.4	68.7	40.9	72.7	79.9
Grupo 2	50	23.0	25.2	31.5	31.9	31.9	29.2	27.0	46.6	30.7	42.8	39.8	39.9	29.9	21.4	46.3	37.5	32.0	28.3	50.0	38.3	31.9	30.9
	100	51.5	50.9	60.7	57.9	57.4	58.0	45.6	66.9	48.1	63.4	64.3	59.6	56.8	44.9	57.7	50.3	50.4	41.6	66.0	52.7	56.1	44.2
Grupo 3	50	53.8	54.7	59.4	58.2	60.9	58.4	49.7	59.9	42.3	58.1	58.5	55.8	42.5	41.3	40.8	41.2	26.6	47.9	39.9	40.8	34.0	41.1
	100	79.4	80.2	81.8	80.1	81.2	81.9	65.7	80.8	52.9	80.4	80.2	76.8	60.4	57.8	49.4	48.4	40.9	68.5	50.7	49.1	46.4	59.8
Grupo 4	50	14.7	14.8	30.3	34.4	19.8	16.9	12.9	16.5	7.3	19.4	30.2	14.4	9.1	10.4	43.5	18.5	16.5	40.7	54.7	30.0	44.9	51.1
	100	32.2	31.3	40.6	44.6	32.9	32.8	24.7	27.6	7.6	27.6	42.9	18.5	10.8	12.9	64.9	27.7	49.8	62.9	73.9	43.3	63.0	69.7
Global	50	34.7	35.8	45.4	45.3	43.7	39.9	34.8	45.4	31.0	44.9	45.5	41.7	33.1	31.8	44.4	32.6	31.5	48.5	47.8	34.6	42.4	48.4
	100	57.1	57.3	66.7	65.9	63.1	61.1	50.8	63.4	41.3	62.7	64.5	58.7	49.5	45.5	58.5	41.7	55.3	68.8	64.1	45.1	61.7	68.0

Tabla 4.57: *Comparación del porcentaje de rechazo para los tres contrastes y todas las profundidades/similaridades. Distribución nula normal.*

			$A(C_{n,p})$							ACC					DS				DSE			
Medida	Distribución	n	PSem	PS	PO	PP	PL_1	PB	PBM	PSem	PS	PO	PP	PL_1	PB	PBM	SO	SP	SB	SBM	SO	SP	SB	SBM
Índice	Normal	50	14.43	13.74	9.86	10.25	10.51	12.53	15.08	9.26	14.35	8.00	8.74	10.49	12.06	12.93	8.57	14.60	15.38	10.08	7.81	13.58	10.97	9.79
		100	12.83	12.33	9.31	10.38	11.22	10.92	15.40	9.94	14.64	9.44	9.28	11.74	11.86	13.60	9.31	15.43	12.11	9.32	8.44	14.06	12.10	9.35
	Uniforme	50	5.78	7.10	7.65	7.78	8.30	6.35	8.43				9.48		6.88	12.43	7.73	9.63	10.15	10.65	13.73	10.55	14.08	14.35
		100	7.53	7.68	8.73	7.78	7.68	7.58	7.80				9.07		8.15	11.63	7.38	10.07	10.90	12.23	12.25	9.88	12.28	12.43
	Exponencial	50	10.02	10.26	7.38	8.45	6.57	7.31	11.71				11.88		11.10	8.76	9.88	11.10	12.98	12.67	9.69	8.48	6.36	6.40
		100	8.93	10.02	8.07	8.10	6.62	7.57	12.26				12.33		11.31	9.38	9.74	10.40	12.95	12.90	7.74	9.12	6.71	6.83
Porcentaje	Normal	50	34.70	35.80	45.40	45.30	43.70	39.90	34.80	45.40	31.00	44.90	45.50	41.70	33.10	31.80	44.40	32.60	31.50	48.50	47.80	34.60	42.40	48.40
		100	57.10	57.30	66.70	65.90	63.10	61.10	50.80	63.40	41.30	62.70	64.50	58.70	49.50	45.50	58.50	41.70	55.30	68.80	64.10	45.10	61.70	68.00
	Uniforme	50	43.05	41.45	41.50	40.95	39.75	42.05	39.10				38.35		39.45	28.15	43.85	33.80	33.75	31.75	20.95	32.95	18.45	17.40
		100	56.80	56.80	58.00	59.75	56.55	56.75	56.35				57.40		54.80	48.60	63.10	52.40	50.05	46.65	48.60	51.45	45.25	41.80
	Exponencial	50	54.40	53.90	59.85	57.80	61.75	60.20	51.60				50.40		52.70	61.05	57.95	56.10	42.70	44.60	55.45	60.85	63.60	64.10
		100	69.50	68.50	71.85	71.65	74.65	72.55	61.20				60.10		62.85	71.85	66.60	67.30	53.80	54.25	72.75	72.05	76.15	76.25

Tabla 4.58: *Comparación del índice y el porcentaje de rechazo para los tres contrastes y todas las profundidades/similaridades. Distribuciones nula normal, uniforme y exponencial.*

		Índice				Porcentaje rechazo			
Distribución	n	$A(C_{n,p})$	ACC	DS	DSE	$A(C_{n,p})$	ACC	DS	DSE
Semiespacial	50	1.85	1.15			34.7	45.4		
	100	1.57	1.43			57.1	63.4		
Simplicial	50	1.61	1.39			35.8	31.0		
	100	1.51	1.49			57.3	41.3		
Oja	50	2.83	2.67	2.44	2.06	45.4	44.9	44.4	47.8
	100	2.76	2.58	2.54	2.11	66.7	62.7	58.5	64.1
Proyecciones	50	2.08	1.89	3.29	2.74	45.3	45.5	32.6	34.6
	100	2.01	1.79	3.50	2.69	65.9	64.5	41.7	45.1
L_1	50	1.57	1.43			43.7	41.7		
	100	1.50	1.50			63.1	58.7		
Bandas	50	2.42	2.50	3.00	2.08	39.9	33.1	31.5	42.4
	100	2.40	2.72	2.53	2.35	61.1	49.5	55.3	61.7
Bandas modificada	50	2.89	2.72	2.40	1.99	34.8	31.8	48.5	48.4
	100	2.83	2.85	2.15	2.17	50.8	45.5	68.8	68.0

Tabla 4.59: *Índice y porcentaje de rechazo para cada similaridad. Distribución nula normal.*

		MA	MC	FA	HZ	GD	NN	Q	$A(C_{n,p})$/PO	ACC/PO	ACC/PP	DS/SO	DS/SBM	DSE/SO	DSE/SBM
Índice	Grupo 1	7.28	8.88	5.81	6.25	6.91	10.72	5.94	8.97	8.56	9.16	6.91	7.03	5.59	7.00
	Grupo 2	8.90	8.80	5.50	4.10	6.60	11.20	5.00	9.40	6.60	8.00	7.00	9.10	5.60	9.20
	Grupo 3	11.75	11.75	5.80	9.55	8.15	10.85	5.65	3.85	3.70	3.50	6.95	7.35	7.50	8.65
	Grupo 4	10.40	13.30	4.50	6.90	11.00	13.20	8.30	9.30	8.80	7.10	3.60	5.30	1.30	2.00
	Global	9.18	10.28	5.58	6.96	7.78	11.17	6.06	7.65	6.97	7.14	6.47	7.17	5.53	7.07
Porcentaje	Grupo 1	49.44	44.75	64.25	58.19	51.56	30.13	60.25	45.75	45.25	44.00	46.31	57.56	49.88	57.56
	Grupo 2	36.80	34.80	52.20	61.40	48.20	22.00	56.20	31.80	42.80	39.80	46.40	28.20	50.00	31.00
	Grupo 3	29.20	29.20	55.00	38.50	43.90	21.20	52.90	59.50	58.20	58.60	40.80	47.90	39.90	41.10
	Grupo 4	22.40	6.20	41.00	27.80	16.60	6.00	23.40	30.20	19.40	30.00	43.40	40.60	54.60	51.20
	Global	38.31	33.69	56.78	48.94	44.11	23.17	52.53	45.47	44.92	45.53	44.39	48.44	47.78	48.42

Tabla 4.60: *Comparación del índice y el porcentaje de rechazo de las mejores combinaciones de contrastes-similaridad/profundidad con otros contrastes existentes. Distribución nula normal.*

Capítulo 5

Conclusiones y futuras líneas de investigación

Resumen

En este capítulo se resumen los resultados principales de la tesis y se enumeran algunas de las posibles futuras líneas de investigación. El desarrollo más importante que se ha introducido en este documento es el de las funciones de similaridad, que preservan la idea de profundidad estadística. Estas funciones de similaridad son la base sobre la que se definen nuevas métricas que, por su forma de medir proximidades, pueden ser de especial interés en Estadística, aplicándolas, por ejemplo, a la construcción de contrastes y a problemas de clasificación en los que, en ocasiones, las distancias empleadas usualmente no son adecuadas. Basado en estas similaridades, se ha propuesto un contraste de bondad de ajuste, que se sitúa al mismo nivel que los mejores contrastes de bondad de ajuste multivariante. Éste, para deteminadas similaridades, puede aplicarse de forma sencilla sobre muestras de dimensión mayor que dos. Además, se ha profundizado en la aplicación de las funciones de profundidad mediante la definición de dos contrastes, uno de dispersión y otro de bondad de ajuste, cuyos resultados están próximos a los obtenidos con el contraste basado en similaridades.

5.1. Conclusiones

A continuación se enumeran las principales conclusiones que se extraen de los resultados recogidos en esta memoria.

La profundidad estadística consiste en la cuantificación de la centralidad de puntos con respecto a una función de distribución. Numerosas funciones se han propuesto en la literatura y cada una de éstas mide la centralidad de una forma particular. Algunas funciones están basadas en medidas de discrepancia o distancias, y otras tienen en cuenta determinados aspectos geométricos de la función respecto a la cual calculan la profundidad. Todas estas funciones tienen en común el objetivo de medir el grado de proximididad entre un punto cualquiera y el centro de una distribución. El primer resultado que se extrae de esta tesis es que es posible realizar extensiones de estas funciones de profundidad, conservando la noción básica de proximidad de cada profundidad, con las que se obtienen otras funciones que permiten la comparación de dos e incluso más puntos simultáneamente. Estas nuevas funciones verifican las propiedades necesarias para que se las considere funciones de similaridad.

Se ha conseguido realizar la adaptación de seis funciones de profundidad conocidas. Para todas estas similaridades se ha llevado a cabo un análisis de las propiedades deseables como extensión de las funciones de profundidad. De estas seis similaridades, tres de ellas (Mahalanobis, proyecciones y Oja) están basadas en funciones que miden la discrepancia o distancia entre puntos. Las otras tres similaridades (simplicial, bandas y bandas modificada) están basadas en aspectos más geométricos de la configuración de los puntos o de la forma de la función de distribución. Para estas tres funciones, se ha realizado un estudio más exhaustivo de sus propiedades, para determinar que, bajo ciertas condiciones, son funciones continuas y sus versiones muestrales convergen a las poblacionales cuando el tamaño muestral aumenta.

A partir de las similaridades ha sido posible, mediante determinadas transformaciones, la definición de nuevas funciones que miden distancias teniendo en cuenta la forma de la función de distribución. De las seis similaridades que se han definido, tan sólo la de Oja no ha sido transformada en distancia. Las similaridades de Mahalanobis y por proyecciones,

se han convertido en distancias de forma trivial, ya que estaban construidas a partir de funciones que ya eran distancias. El resultado más importante con respecto a las distancias por profundidad se encuentra en la obtención de las distancias simplicial, por bandas y por bandas modificadas, ya que son funciones más novedosas y poseen además una mayor capacidad de adaptación que las otras a la forma de la función de distribución, lo que las hace más atractivas para su aplicación en problemas multivariantes.

Las similaridades y las distancias por profundidad se han aplicado en el análisis de conglomerados, obteniéndose mejoras con respecto a la distancia euclídea que es la más utilizada en este tipo de problemas. Más concretamente, las similaridades han sido aplicadas en el análisis de conglomerados jerárquico, para el que, tomando como criterio de agrupación el método de Ward, todas las similaridades excepto la de proyecciones, han obtenido errores de clasificación menores que los obtenidos por medio de la distancia euclídea. Siendo las tres similaridades con mejores tasas de error la de bandas, simplicial y bandas modificada. Por su parte, las distancias se han utilizado también para el análisis de conglomerados, pero sobre una modificación del algoritmo de k-medias. Del estudio de los resultados con este algoritmo modificado se ha concluido, por un lado, que las distancias por profundidad son muy sensibles a la elección de los centros iniciales sobre los que se ejecuta el algoritmo y, por otro lado, que el número de mínimos locales en los que el algoritmo puede detenerse parece ser superior al del algoritmo de k-medias. Y también, que una vez eliminado el efecto de los centros iniciales, las tasas de errores en la agrupación para determinadas distancias son mucho menores que las obtenidas con la distancia euclídea y con el algoritmo de k-medias. De nuevo, las distancias que obtienen mejores resultados en las agrupaciones son la de bandas, simplicial y de bandas modificada.

Sobre las funciones de profundidad, se han diseñado dos contrastes de hipótesis. En el primero, el objetivo es determinar si la dispersión de un conjunto de datos es igual a la dispersión de una determinada función de distribución. Este contraste se basa en la curva de escala, que mide la evolución del volumen desde los puntos más profundos hasta los más externos. Del estudio de la potencia, para contrastar si la dispersión es

igual que la de la distribución normal multivariante, se concluye que las profundidades que en promedio poseen una potencia más elevada son la de Oja, la L_1 y la de bandas. El segundo contraste es un contraste de bondad de ajuste. Éste estudia cuánto se parecen las regiones centrales de muestra y distribución nula. Las discrepancias se recogen en dos curvas: una que muestra la concordancia de la muestra con la distribución y otra que recoge la discrepancia en sentido inverso. Si la muestra procede de la distribución nula, ambas curvas son iguales e iguales a su vez a la recta $y = x$. Tomando como hipótesis nula la distribución normal multivariante, del estudio de la potencia del contraste se tiene que las profundidades para las que el contraste es más potente son la de Oja y la de proyecciones.

El último contraste que se ha introducido en esta memoria es un contraste de bondad de ajuste basado en las similaridades por profundidad. Las similaridades por profundidad entre dos puntos se calculan con respecto a una función de distribución. En este contraste se obtienen las similaridades entre puntos de la muestra con respecto a la función de distribución empírica y con respecto a la distribución nula y se comparan. Si la muestra sigue la distribución nula las diferencias son pequeñas. Las similaridades que obtienen mayor potencia son la de Oja y la de bandas modificada.

Los resultados obtenidos en los tres contrastes para la hipótesis de distribución normal, se comparan con los obtenidos para otros contrastes de la literatura obteniédose resultados competitivos. Sobre todo para el contraste basado en la similaridad de Oja que ocupa la primera posición en esta comparación si se tiene en cuenta un índice calculado a partir de la posición sobre todas las distribuciones alternativas y la sexta si se tiene en cuenta la potencia media. El basado en la similaridad por bandas modificada también está entre los primeros y por delante de contrastes de normalidad como los de asimetría y curtosis de Mardia. Para la distribución nula uniforme el contraste que mejores resultados obtiene es el de la curva de volumen, mientras que, para la distribución nula exponencial, lo es el basado en la similaridad por bandas y por bandas modificada. Esto sugiere la posibilidad de que para variables aleatorias que tomen valores en regiones acotadas, sea mejor emplear el contraste de la curva de escala y, en caso contrario, el basado en la similaridad.

Como última conclusión y valorando de forma global los resultados de las aplicaciones, se tiene que para problemas de clasificación, las similaridades y distancias simplicial, de bandas y de bandas modificada funcionan mucho mejor que el resto. Mientras que, para la determinación de la bondad de ajuste, las mejores son con diferencia la profundidad y similaridad de Oja, seguidas de la similaridad por bandas modificada y de las funciones de profundidad por proyecciones y L_1.

5.2. Futuras líneas de investigación

La investigación desarrollada en esta tesis doctoral sugiere varias líneas de posible investigación futura que se presentan a continuación.

5.2.1. Extensión de otras funciones de profundidad

Aparte de las seis profundidades extendidas en esta tesis, existen numerosas funciones de profundidad en la literatura. Una línea de posible investigación consiste en la extensión y análisis de las propiedades para otras funciones de profundidad. Por ejemplo, se pueden extender la profundidad del zonoide y la profundidad L_1, la cual puede ser de una gran utilidad en problemas de clasificación en alta dimensión debido al bajo coste computacional necesario para su cálculo.

5.2.2. Refinamiento de los centros iniciales en clasificación no supervisada

Debido al problema de sensibilidad al emplear las distancias en el algoritmo de k-medias, se hace necesaria la aplicación de algún método de refinamiento de centros iniciales. Por lo tanto debe realizarse un análisis exhaustivo de los métodos de centros iniciales para determinar cuál de ellos produce resultados más fiables para las distancias por profundidad. También es necesario analizar los criterios de convergencia del algoritmo para adecuarlos a la naturaleza de las distancias, con el objetivo de reducir el número de mínimos locales.

5.2.3. Aplicación de las distancias por profundidad en clasificación supervisada

La clasificación supervisada no ha sido abordada en esta memoria. Se propone introducir las distancias basadas en profundidad de dos formas. La primera, como matriz de distancias entre puntos, calculada con respecto a una mixtura entre la distribución empírica de toda la muestra y una distribución continua. Y la segunda, a través de tantas matrices como grupos haya en la muestra y que se obtengan con respecto a la mixtura entre la distribución empírica de los componentes del grupo y una distribución continua. En el segundo caso habría que analizar además las escalas de las matrices de cada grupo para hacer que sean comparables y que, por tanto, las asignaciones no dependan tanto de estas escalas.

5.2.4. Similaridades y distancias en grupos

En todas las aplicaciones de esta memoria se miden las similaridades y las distancias entre pares de puntos. En problemas de clasificación y agrupamiento resulta de utilidad el poder realizar comparaciones de más de dos puntos simultáneamente para obtener una medida global de las distancias o similitudes dentro de un grupo. Es conveniente estudiar su introducción tanto como criterio de aglomeración en el análisis de conglomerados jerárquico como representación de la heterogeneidad global dentro de un grupo en el algoritmo de k-medias o modificaciones suyas. Las similaridades de Oja, simplicial, por bandas y por bandas modificadas permiten de forma inmediata la inclusión de más puntos.

5.2.5. Modificaciones del contraste de escala

En el contraste de la curva de escala, el estadístico del contraste es el área entre la curva muestral y la curva esperada cuando la hipótesis nula es cierta. Este estadístico no tiene en cuenta que la variabilidad al principio de la curva es menor que al final. Se pretende estudiar si mejora la potencia del contraste al cambiar la metodología como se expone a continuación. Llevar a cabo la simulación de un número elevado B de muestras

bajo la hipótesis nula, para las que se obtiene la curva de escala. Posteriormente, calcular la profundidad de la curva de escala de la muestra que se desea contrastar con respecto a la distribución empírica de las B curvas simuladas y rechazar cuando dicha profundidad sea menor que el percentil de nivel α calculado a partir de las profundidades de las B curvas de escala simuladas.

REFERENCIAS

Anderson, T. W. y Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769.

Arcones, M. A., Chen, Z. Q., y Giné, E. (1994). Estimators related to U-processes with applications to multivariate medians - asymptotic normality. *Annals of Statistics*, 22(3):1460–1477.

Arcones, M. A. y Giné, E. (1993). Limit-theorems for U-processes. *Annals of Probability*, 21(3):1494–1542.

Bai, Z. D. y He, X. M. (1999). Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *Annals of Statistics*, 27(5):1616–1637.

Baringhaus, L., Danschke, R., y Henze, N. (1989). Recent and classical tests for normality - a comparative-study. *Communications in Statistics-Simulation and Computation*, 18(1):363–379.

Barnett, V. (1976). Ordering of multivariate data. *Journal of the Royal Statistical Society Series A-Statistics in Society*, 139:318–354.

Bartoszynski, R., Pearl, D. K., y Lawrence, J. (1997). A multidimensional goodness-of-fit test based on interpoint distances. *Journal of the American Statistical Association*, 92(438):577–586.

Bickel, P. J. y Lehmann, E. L. (1975). Descriptive statistics for nonparametric models 2: Location. *Annals of Statistics*, 3(5):1045–1069.

Bickel, P. J. y Lehmann, E. L. (1976). Descriptive statistics for nonparametric models 3: Dispersion. *Annals of Statistics*, 4(6):1139–1158.

Csörgo, S. y Faraway, J. J. (1996). The exact and asymptotic distributions of Cramer-von Mises statistics. *Journal of the Royal Statistical Society Series B-Methodological*, 58(1):221–234.

Cui, H. J. y Cheng, P. (1996). The p-values of testing for multinormality based on the PP skewness and kurtosis. *Progress in Natural Science*, 6(3):277–283.

D'Agostino, R. B. (1971). Omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2):341–&.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises tests. *Annals of Mathematical Statistics*, 28(4):823–838.

Donoho, D. L. y Gasko, M. (1992). Breakdown properties of location estimates based on half-space depth and projected outlyingness. *Annals of Statistics*, 20(4):1803–1827.

Dümbgen, L. (1992). Limit-theorems for the simplicial depth. *Statistics & Probability Letters*, 14(2):119–128.

Epps, T. W. y Pulley, L. B. (1983). A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726.

Fattorini, L. (1986). Remarks on the use of the Shapiro-Wilks statistic for testing multivariate normality. *Statistica*, 46(2):209–217.

Fraiman, R. y Meloche, J. (1999). Multivariate L-estimation. *Test*, 8(2):255–289.

Hall, P. y Welsh, A. H. (1983). A test for normality based on the empirical characteristic function. *Biometrika*, 70(2):485–489.

Hartigan, J. (1975). *Clustering Algorithms*. John Wiley and Sons, New York.

He, X. M. y Wang, G. (1997). Convergence of depth contours for multivariate datasets. *Annals of Statistics*, 25(2):495–504.

Henze, N. y Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, 19(10):3595–3617.

Hettmansperger, T. P., Nyblom, J., y Oja, H. (1994). Affine invariant multivariate one-sample sign tests. *Journal of the Royal Statistical Society Series B-Methodological*, 56(1):221–234.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–235.

Huber, P. J. (1972). 1972 Wald lecture - robust statistics - review. *Annals of Mathematical Statistics*, 43(4):1041–1067.

Hueter, I. (1999). Limit theorems for the convex hull of random points in higher dimensions. *Transactions of the American Mathematical Society*, 351(11):43374363.

Jarque, C. M. y Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 55(2):163–172.

Jornsten, R. (2004). Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis*, 90(1):67–89.

Jornsten, R., Vardi, Y., y Zhang, C. (2002). A robust clustering method and visualization tool based on data depth. *Statistical data analysis based on the L_1-norm and related methods. Birkhauser 2002, Statistics for industry and technology. Y. Dodge editor.*

Justel, A., Pena, D., y Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters*, 35(3):251–259.

Kaufman, L. y Rousseeuw, P. J. (1986). *Pattern Recognition in Practice II*, pages 425–437. Elsevier/North-Holland, Amsterdam.

Kaufman, L. y Rousseeuw, P. J. (1987). Clustering by means of medoids. *Statistical Data Analysis based on the L_1-Norm*, pages 405–416.

Kaufman, L. y Rousseeuw, P. J. (1990). *Finding groups in data*. Wiley Series in Probability and Mathematical Statistics.

Koshevoy, G. y Mosler, K. (1997). Zonoid trimming for multivariate distributions. *Annals of Statistics*, 25(5):1998–2017.

Koziol, J. A. (1986). Assessing multivariate normality - a compendium. *Communications in Statistics-Theory and Methods*, 15(9):2763–2783.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *Annals of Statistics*, 18(1):405–414.

Liu, R. Y., Parelius, J. M., y Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics*, 27(3):783–840.

Liu, R. Y. y Singh, K. (1992). Ordering directional-data - concepts of data depth on circles and spheres. *Annals of Statistics*, 20(3):1468–1484.

Liu, R. Y. y Singh, K. (1993). A quality index based on data depth and multivariate rank-tests. *Journal of the American Statistical Association*, 88(421):252–260.

Liu, R. Y. y Singh, K. (1997). Notions of limiting p-values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92(437):266–277.

López-Pintado, S. y Romo, J. (2007). Depth-based inference for functional data. *Computational Statistics & Data Analysis*, 51(10):4957–4968.

López-Pintado, S. y Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.

Lorenz, M. (1905). Methods of measuring the concentration of wealth. *Journal of the American Statistical Association*, 9:209–219.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. En *Proceedings National Institute of Science, India*, volume 2, pages 49–55.

Malkovic, J. y Afifi, A. A. (1973). Tests for multivariate normality. *Journal of the American Statistical Association*, 68(341):176–179.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530.

Massé, J. (2002). Asymptotics for the Tukey median. *Journal of Multivariate Analyisis*, 81(2):286–300.

Massé, J. C. y Theodorescu, R. (1994). Half-plane trimming for bivariate distributions. *Journal of Multivariate Analysis*, 48(2):188–202.

Miller, K., Ramaswami, S., Rousseeuw, P., Sellares, J. A., Souvaine, D., Streinu, I., y Struyf, A. (2003). Efficient computation of location depth contours by methods of computational geometry. *Statistics and Computing*, 13(2):153–162.

Mizera, I. y Volauf, M. (2002). Continuity of halfspace depth contours and maximum depth estimators: Diagnostics of depth-related methods. *Journal of Multivariate Analysis*, 83(2):365–388.

Nolan, D. (1992). Asymptotics for multivariate trimming. *Stochastic Processes and Their Applications*, 42(1):157–169.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1:327–332.

Pearson, E. S., D'Agostino, R. B., y Bowman, K. O. (1977). Tests for departure from normality - comparison of powers. *Biometrika*, 64(2):231–246.

Quiroz, A. J. y Dudley, R. M. (1991). Some new tests for multivariate normality. *Probability Theory and Related Fields*, 87(4):521–546.

Romeu, J. L. y Ozturk, A. (1993). A comparative-study of goodness-of-fit tests for multivariate normality. *Journal of Multivariate Analysis*, 46(2):309–334.

Rousseeuw, P. J. y Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, 94(446):388–402.

Rousseeuw, P. J. y Ruts, I. (1996). Bivariate location depth. *Applied Statistics-Journal of the Royal Statistical Society Series C*, 45(4):516–526.

Rousseeuw, P. J., Ruts, I., y Tukey, J. W. (1999). The bagplot: A bivariate boxplot. *American Statistician*, 53(4):382–387.

Rousseeuw, P. J. y Struyf, A. (1998). Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 8(3):193–203.

Royston, J. P. (1982a). An extension of Shapiro and Wilk-W test for normality to large samples. *Applied Statistics-Journal of the Royal Statistical Society Series C*, 31(2):115–124.

Royston, J. P. (1982b). The W test for normality. *Applied Statistics-Journal of the Royal Statistical Society Series C*, 31(2):176–180.

Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk-W. *Applied Statistics-Journal of the Royal Statistical Society Series C*, 32(2):121–133.

Shapiro, S. S. y Wilk, M. B. (1965). An analysis-of-variance test for normality. *Biometrika*, 52:591–611.

Shapiro, S. S., Wilk, M. B., y Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Associaction.*, 63:1343–1372.

Sinclair, C. D., Spurr, B. D., y Ahmad, M. I. (1990). Modified Anderson Darling test. *Communications in Statistics-Theory and Methods*, 19(10):3677–3686.

Singh, K. (1991). Majority depth. Unpublished manuscript.

Stephens, M. A. (1974). Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737.

Székely, G. J. y Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80.

Tukey, J. W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians, Vancouver*, pages 523–531.

Van Aelst, S. y Rousseeuw, P. J. (2000). Robustness of deepest regression. *Journal of Multivariate Analysis*, 73(1):82–106.

Van Aelst, S., Rousseeuw, P. J., Hubert, M., y Struyf, A. (2002). The deepest regression method. *Journal of Multivariate Analysis*, 81(1):138–166.

Vardi, Y. y Zhang, C. H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America*, 97(4):1423–1426.

Vasicek, O. (1976). Test for normality based on sample entropy. *Journal of the Royal Statistical Society Series B-Methodological*, 38(1):54–59.

Watson, G. S. (1957). The χ^2 goodness-of-fit test for normal distributions. *Biometrika*, 44(3-4):336–348.

Watson, G. S. (1958). On chi-square goodness-of-fit tests for continuous distributions. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 20(1):44–72.

Watson, G. S. (1959). Some recent results in chi-square goodness-of-fit tests. *Biometrics*, 15(3):440–468.

Yeh, A. B. y Singh, K. (1997). Balanced confidence regions based on Tukey's depth and the bootstrap. *Journal of the Royal Statistical Society Series B-Methodological*, 59(3):639–652.

Zhou, S. y Jammalamadaka, S. R. (1993). Goodness of fit in multidimensions based on nearest-neighbour distances. *Nonparametric Statistics*, 2(3):271–284.

Zhu, L. X., Fang, K. T., y Bhatti, M. I. (1997). On estimated projection pursuit-type Cramer-von Mises statistics. *Journal of Multivariate Analysis*, 63(1):1–14.

Zhu, L. X., Wong, H. L., y Fang, K. T. (1995). A test for multivariate normality based on sample entropy and projection pursuit. *Journal of Statistical Planning and Inference*, 45(3):373–385.

Zuo, Y. J. (2003). Projection-based depth functions and associated medians. *Annals of Statistics*, 31(5):1460–1490.

Zuo, Y. J. y Serfling, R. (2000a). General notions of statistical depth function. *Annals of Statistics*, 28(2):461–482.

Zuo, Y. J. y Serfling, R. (2000b). Structural properties and convergence results for contours of sample statistical depth functions. *Annals of Statistics*, 28(2):483–499.